UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF MECHANICAL ENGINEERING

Postgraduate Thesis

APPROACHES FOR OPTIMAL LOCATION OF EMERGENCY RESPONSE VEHICLES

by

ATHANASIOS PANTIDIS

with an Electrical & Computer Engineering Diploma D.U.TH., 2002

Submitted for the completion of a part of requirements which lead to

Postgraduate Specialization Diploma

2009

Approved by the three Member Examination Committee:

Major Advisor	Professor Ziliaskopoulos Thanassis Department of Mechanical Engineering, University of Thessaly
Member (Co-advisor)	Professor Liberopoulos George Department of Mechanical Engineering, University of Thessaly
Member (Co-advasor)	Lecturer Kozanidis George Department of Mechanical Engineering, University of Thessaly

Acknowledgments

I would like to thank Professor Ziliaskopoulos Thanassis for being a great advisor. My thanks to Professor Liberopoulos George and Kozanidis George for being members of my Examination Committee, for reading this thesis and for offering valuable suggestions. I would also like to thank Vaggelis Katsaros, Ph.D Student, for his collaboration during this thesis process.

Athanasios Pantidis

APPROACHES FOR OPTIMAL LOCATION OF EMERGENCY RESPONSE VEHICLES

PANTIDIS ATHANASIOS

Democritus University of Thrace, Department of Electrical & Computer Engineering, 2002

Advisor: Professor Ziliaskopoulos Thanassis in Optimization of Production/Transportation Systems

Abstract

Police, fire and emergency medical systems are all concerned with improving public safety, and share the common objective of responding to citizen calls for assistance as quickly as possible to reduce loss of life and injury.

Optimization of emergency response vehicles location is a research area which is concerned with the location of one or more vehicles so as to satisfy objective function requirements such as providing fast and reliable service to customers. The most important decision facing any emergency response service is how many emergency vehicles to have, and on which site to locate them.

A vast literature has developed out of the significant research interest in meeting this challenge. The literature review is separated into three sections depending on the objective function of the location models: Covering models, P-median models, and Center models.

In the next chapter 3, we describe characteristics and performance criteria of emergency response services. The assumption is that if calls are answered and serviced quickly, then this will lead to customer satisfaction and compliance to regulatory standards for response time performance. The decision-maker is confronted with the elements of time and distance simultaneously. The time taken to get to an incident is necessarily dependent upon the distance to be travelled and the conditions experienced during the journey. Timeliness, cost minimization, coverage equity maximization and labor equity maximization are the most important objectives of emergency service systems. In this thesis we also, focused on the description of some methods to estimate travel distance and travel time. A crucial issue in locating emergency response vehicles is data availability. Collection and analysis of the available data point out one of the main problems of the system.

Mathematical models may be very useful in dealing with emergency response vehicle location. In chapter 4, location models are classified according to their objectives, constraints, solutions, and other attributes.

There has been an important evolution in the development of emergency vehicles location and relocation models over the past years. In this thesis, we attempted to provide an

overview of emergency vehicles location models dedicated to capturing the complex time and uncertainty characteristics of most real-world problems.

Chapter 5 concerns an elaborate description of the basic emergency response vehicles location models, mostly, in discrete space or networks, that are related to the public sector, such as ambulances, fire vehicles, police units. Static and deterministic location models assume that the nearest unit to a call for service is always available. Dynamic models can be used to periodically update emergency vehicles positions throughout the day. Probabilistic models deal with the stochastic nature of real-world systems. In these systems, models capture the stochastic aspects of facility location through explicit consideration of the probability distributions associated with modeled random quantities. Parameters, such as travel times, the location of clients, demand and the availability of servers are treated as random variables. The objective is to determine robust server/facility locations that optimize a given utility function, for a range of values of the parameters under consideration.

Finally, in chapter 6 we present two applications of P- Median and Hypercube models. The solution of Hypercube model is the state probabilities and associated system performance measures such as workloads. As far as it concerns the P-Median model, the aim is to locate a fixed number of vehicles so as to minimize the weighted travel time of the system. In the end, we solve P-Median model for fixed number of servers and we implement the hypercube model using the assignment resulted from P-Median problem.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

A fire engine or ambulance speeding to the scene of an emergency, or a police car patrolling city streets are common images of daily life. In the mid-1960's operations researchers began studying the deployment of these emergency services. Police, fire and emergency medical systems are all concerned with improving public safety, and share the common objective of responding to citizen calls for assistance as quickly as possible to reduce loss of life and injury.

Optimization of emergency response facilities (vehicles) location is a research area in operations research concerned with the location of one or more facilities so as to satisfy objective function requirements such as providing fast and reliable service to customers. As populations shift, the need to relocate, expand, and adapt facilities ensures the evolution of new planning challenges.

Before a facility can be constructed or located, good locations must be identified, appropriate facility capacity specifications must be determined, and large amounts of capital must be allocated. While the objectives driving a facility location decision depend on the firm or government agency, the high costs associated with this process make almost any location project a long-term investment. Thus, facilities which are located today are expected to remain in operation for an extended time. Environmental changes during the facility's lifetime can drastically alter the appeal of a particular site, turning today's optimal location into

tomorrow's investment misstep. Determining the best locations for new facilities is thus an important strategic challenge.

Perhaps the most important decision facing any emergency response service is how many fire vehicles or ambulances to have, and on which site to locate them. It is appreciated that the optimum solution is the one which minimizes the sum of losses providing fast and reliable service to customers.

Chapter 2 Literature Review

Determining the optimal location of emergency vehicles such as ambulances, fire vehicles etc, is an important strategic and operational consideration. Emergency response facilities (vehicles) should be located in such a way as to ensure an adequate coverage and a quick response time.

A vast literature has developed out of the significant research interest in meeting this challenge. Also, unprecedented growth in computer power and in algorithmic sophistication has contributed to a significant evolution in location models. This review is separated into three sections depending on the objective function of the location models: Covering models, P-median models, and Center models.

2.1 Covering Models for Emergency Services

Covering models are the most widespread location models for formulating the emergency facility location problems. The objective of covering models is to provide "coverage" to demand points. A demand point is considered as covered only if a facility is available to service the demand point within a distance limit. Covering problems are divided into two major parts: the location set covering problem (LSCP) and the maximal covering location problem (MCLP).

LSCP is an earlier statement of the emergency facility location problem by Toregas et al. in 1971 [39] and it aims to locate the least number of facilities that are required to cover all

demand points. Since all the demand points need to be covered in LSCP, regardless of their population, remoteness, and demand quantity, the resources required for facilities could be excessive. Recognizing this problem, Church and ReVelle in 1974 [6] developed the MCLP model that does not require full coverage to all demand points. Instead, the model seeks the maximal coverage with a given number of facilities. The MCLP, and different variants of it, have been extensively used to solve various emergency service location problems. A notable example is the work of Eaton et al. in 1985 [11] that used MCLP to plan the emergency medical service in Austin, Texas. The solution gives a reduced average emergency response time even with increased calls for service. Neither LSCM nor MCLP recognizes the fact that on occasions vehicles of several types may be dispatched to the scene of an incident. Also, even if only one vehicle type is used, solving MCLP alone may not provide a sufficiently robust location plan. One of the first models developed to handle several vehicle types is the tandem equipment allocation model, or TEAM [37]. Schilling et al. in 1979 generalized the MCLP model to locate emergency fire-fighting servers and depots in the city of Baltimore. In their model, known as FLEET (Facility Location and Equipment Emplacement Technique), two different types of servers need to be located simultaneously. A demand point is regarded as "covered" only if both servers are located within a specified distance.

The preceding models do not consider the system congestion and unavailability of the facilities. Many covering models have also been developed to address the possible congestion condition by providing redundant or back-up coverage. Daskin and Stern in 1981 [9] formulated a hierarchical objective LSCP for emergency medical service in order to find the minimum number of vehicles that are required to cover all demand areas while simultaneously maximizing the multiple coverage. Hogan and ReVelle in 1986 [19] developed MCLP models for emergency service that has a secondary "backup-coverage"

objective. The models ensure that a second (backup) facility could be available to service a demand area in case that the first facility is unavailable to provide services. The backup coverage models have been popularly called as BACOP1 (Backup Coverage Problem 1). Since the models of BACOP1 require each demand point to have first coverage which is not necessary for many location problems, Hogan and ReVelle in 1986 [19] further formulated the BACOP2 model which is able to respectively maximize the population that achieve first and second coverage.

Gendreau et al. in 2001 [14], developed a model that considers the objective of maximizing double coverage of demand. The constraints include: the number of vehicles at each site, the moving of the same vehicle repeatedly, long travel trips, and round trips between two sites.

Research on emergency service covering models has also been extended to incorporate the stochastic and probabilistic characteristics of emergency situations so as to capture the complexity and uncertainty of these problems. There are several approaches to model stochastic emergency service covering problems. Daskin in 1983 [12] used an estimated parameter (q) to represent the probability that at least one server is free to serve the requests from any demand point. He formulated the Maximum Expected Covering Location Problem (MEXCLP) to place P facilities on a network with the goal to maximize the expected value of population coverage. ReVelle and Hogan in 1986 [34] later enhanced the MEXCLP and proposed the Probabilistic Location Set Covering Problem (PLSCP). In the PLSCP, an average server busy fraction (q_i) and a service reliability factor (α) are defined for the demand points. Then the locations of facilities are determined such that the probability of service being available within a specified distance is maximized. The MEXCLP and PLSCP later were further modified to tackle other emergency service location problems by ReVelle and

Hogan in 1989 [35] and Repede and Bernardo in 1994 [32]. Repede and Bernardo in 1994 [32] extended Daskin's maximal expected covering model to allow different location sets at different times of the week.

A cornerstone in location theory is the development and application of the queuing approach in solving emergency service location problems. The most well known queuing models are the hypercube and approximated hypercube by Larson [23, 24], which consider the congestions of the system by calculating the steady-state busy fractions of servers on a network. The hypercube model can be used to evaluate a wide variety of output performance such as vehicle utilization, average travel time, inter-district service performance, etc. Marianov and ReVelle in 1996 [26] created a realistic location model for emergency systems based on results from queuing theory. In their model, the travel times or distances along arcs of the network are considered as random variables. The goal is to place limited numbers of emergency vehicles, such as ambulances, in a way as to maximize the calls for service.

2.2 P-Median Models for Emergency Services

Another important way to measure the effectiveness of facility location is by evaluating the average total distance between the demand points and the facilities. When the average total distance decreases, the accessibility and effectiveness of the facilities increases. The P-median problem, introduced by Hakimi in 1964 [16], takes this measure into account and is defined as: determine the location of P facilities so as to minimize the average (total) distance between demands and facilities.

Since its formulation the P-median model has been enhanced and applied to a wide range of emergency facility location problems. One major application of the P-median models is to dispatch emergency response units such as ambulances during emergency incidents.

Carson and Batta in 1990 [5] proposed a P-median model to find the dynamic ambulance positioning strategy for campus emergency service. The model uses scenarios to represent the demand conditions at different times. The ambulances are relocated in different scenarios in order to minimize the average response time to the service calls. Berlin et al. in 1976 [2], investigated two P-median problems to locate hospitals and ambulances. The first problem has a major attention to patient needs and seeks to minimize the average distance from the hospitals to the demand points and the average ambulance response time from ambulance bases to demand points. In the second problem, a new objective is added in order to improve the performance of the system by minimizing the average distance from ambulance bases to hospitals. Mandell in 1998 [4], developed a P-median model and used priority dispatching to optimally locate emergency units for a tiered EMS system that consists of advanced life-support (ALS) units and basic life-support (BLS) units. The model can also be used to examine other system parameters including the balance between ALS and BLS units, and different dispatch rules.

Uncertainties have also been considered in many P-median models. Mirchandani in 1980 [29], examined a P-median problem to locate fire-fighting emergency units with consideration of stochastic travel characteristics and demand patterns. He took into account the situations that a facility may not be available to serve a demand and used a Markov process to create a system in which the states were specified according to demand distribution, service and travel time, and server availability. Serra and Marianov in 1996 [38], implemented a P-median model and introduced the concept of regret and minmax objectives when locating fire station for emergency services in Barcelona. The authors addressed in their model the issue of locating facilities when there are uncertainties in demand, travel time or

distance. In addition, the model uses scenarios to incorporate the variation of uncertainties and seeks to give a compromise solution by minimizing the maximum regret over the scenarios.

P-median models have also been extended to solve emergency service location problems in a queuing theory context. An example is the stochastic queue median model due to Berman et al. in 1985 [3]. This model seeks to optimally dispatch mobile servers such as emergency response units to demand points and locate the facilities so as to minimize average cost of response.

2.3 Center Models for Emergency Services

In contrast to the P-median models which concentrate on optimizing the overall (or average) performance of the system, the P-center model attempts to minimize the worst performance of the system and thus addresses situations in which service inequity is more important than average system performance. In location literature, the P-center model is also referred to as the minimax model since it minimizes the maximum distance between any demand point and its nearest facility. The P-center model considers a demand point is served by its nearest facility and therefore full coverage to all demand points is always achieved. However, unlike the full coverage in the set covering models, which may lead to excessive number of facilities, the full coverage in the P-center model requires only a limited number (P) of facilities.

The problem asks for the center of a circle that has the smallest radius to cover all desired destinations. In order to locate a given number of emergency facilities along a road network, Garfinkel et al. in 1977 [15], examined the fundamental properties of the P-center problem. He modeled the P-center problem using integer programming and the problem was successfully solved by using a binary search technique and a combination of exact tests and

heuristics. ReVelle and Hogan in 1989 [35], formulated a P-center problem to locate facilities so as to minimize the maximum distance within which emergency service is available with α reliability. System congestion is considered and a derived server busy probability is used to constrain the service reliability level that must be satisfied for all demands.

Stochastic P-center models have also been formulated for emergency response location problems. For example, Hochbaum and Pathria in 1998 [18] considered the emergency facility location problem that must minimize the maximum distance on the network across all time periods. The cost and distance between locations vary in each discrete time period.

Chapter 3 Characteristics of Emergency Response Service Problems and Performance Criteria

The primary objective of emergency response services is to get the appropriate equipment to calls in a safe and timely fashion. The assumption is that if calls are answered and serviced quickly, then this will lead to customer satisfaction and compliance to regulatory standards for response time performance. The steps of the standard emergency call process are:

1. The call (demand) comes to the system via phone or some other mechanism.
2. The severity of the call is estimated.
3. The dispatcher evaluates the system status and determines the appropriate vehicle or vehicles to send to the scene.
4. Upon arriving the scene, service is provided.
5. The vehicle(s) may or may not provide transport to a hospital.
6. After completion of service (and transport) the vehicle goes into an idle state and returns to a predetermined location to await another call.

The decisions of dispatching and vehicle location are critical factors in system success. If one cannot do both of these well, there will be inefficiencies in the system. The issue of

timeliness is the primary objective that is used in operations research models. We can make the following assumptions:

- There is a standard time, T, such that if the first vehicle arrives on scene within T minutes, then the call service is deemed a success. The specific value of T may vary with the type of call as more serious calls have lower T values.
- The area is partitioned into zones. These zones may take on any shape, but all calls from a zone originate in the population center. All travel to and from the zone is measured from the zone center point. Data is collected and aggregated at the zone level.

There are many ways that timeliness is measured. For example, one can operate to minimize the total or average time/distance to serve all calls. When a call for assistance arrives, fire fighters or staff of ambulance, are automatically confronted with several questions to be answered in a few seconds. These questions include [31]:

- ✓ The location of the incident i.e. fire, car accident (elements, demand, distance)
- ✓ The fastest route to the location of the incident
- ✓ If the shortest distance could be taken (elements of distance and time, where the shortest distance or the minimum time could mean the most congested route)
- ✓ If the shortest distance is not ideal, what is the alternative route should be taken (elements of time and distance, where maximum distance or time could be the least congested route)
- ✓ If average distance means a longer time, what is the minimum maximum distance to the incident (elements of distance and time, where minimum maximum time may not necessarily reflect average distance)

✓ The shortest route should be taken even though it usually reflects maximum time (elements of distance and time, where the major assumption is that the public will cooperate fully in allowing smooth passage for vehicles).

So, the decision-maker is confronted with the elements of time and distance simultaneously. The time taken to get to an incident is necessarily dependent upon the distance to be travelled and the conditions experienced during the journey.

A second way that timeliness is measured is the minimization of the maximum travel time/distance to any single call (ensures that no demand point is too far from vehicle location). It is used to reflect the worst scenario case associated with bad conditions experienced during the journey.

Another way that timeliness is measured is the maximization of area coverage. In this case, we cover as many zones in the area as possible within T minutes of travel.

A forth way that timeliness is measured is the maximization of call coverage so, we cover as many calls in the area as possible within T minutes of travel.

One criterion for judging the performance of vehicle location is the speed at which the system reacts when a call is logged. As a result, the initial spatial allocation of vehicles, influences powerfully the efficiency of the response. To decide on a spatial allocation, requires that a number of issues be addressed. For example, how many vehicles are needed. Usually, there is a priori budget constraints, as a result, the minimum number of vehicles might be sought that would achieve "coverage" of all possible customers. This solution might or might not coincide with the solution that minimizes the cost of the system. Otherwise, the goal might be to "cover" the maximum population with a "good quality of coverage". Thus, response times are determined by an explicit or implicit compromise between the cost of

service and the cost of failure (damage caused to customers by arriving late). It is important that, shorter response times impose greater resource requirements on the system, which in turn increases its cost.

Besides timeliness, another objective of emergency service systems is cost minimization. Cost is primarily a function of the amount of labor (man-hours) needed to staff the unit-hours used per year and the number of vehicles that must be purchased, supported, and serviced.

Another objective of emergency service deployment systems is the coverage equity maximization. Areas are not equal. Depending on several factors, some areas experience higher rates of incidents i.e. fire or car accidents. The system manager must balance area performance against the performance in a smaller group of zones. For example, it may not be acceptable to have zones that are poorly served while having the area at a reasonable level and some zones that are extremely well served. By changing decisions, more equitable systems can be designed.

A forth objective of emergency service deployment systems is the labor equity maximization. It is important for the system manager to balance the workload for all employees in the system. This reduces employee burnout and hard feelings. [1, 28, 33]

3.1 Modeling Issues - Granularity, Data Requirements and Validity

It is often difficult and expensive to experiment with an actual system. Mistakes are costly both in money and in potential mortality. Collecting data to verify a good system might take months of data collection. Instead of experimenting on the actual system, operations research professionals generally build models of systems that can be implemented and experimented with on a computer. System errors can be found on the model before they are

implemented on the actual system. It is generally well worth the cost of building the model, collecting data, and running the model as opposed to trying to experiment on the actual system.

When using a model to help make decisions, there is significant work that must be done before any analysis can begin. First one must structure the granularity of the model and define zones. Next, one gathers demand, service, and travel time data based on that structure. Third, the model is implemented usually in software. Finally, one validates the model to convince the decision maker that model output has some correlation with the output of the actual system.

3.1.1 Granularity of the Zone Structure

The zone structure is often formed based on the convenience of the model builder or the data collection system. Since most urban and suburban fire service systems or emergency medical systems have tens of thousands of calls per year, it is impossible to model down to the call level. Instead, all calls in a "small area" are aggregated to a single zone. Let's consider the aggregation in Figure 3-1. Here, we have 9 calls, one in each of the nine address blocks. Instead of using these individual locations, we aggregate all calls to the center of the blocks and this is our zone.

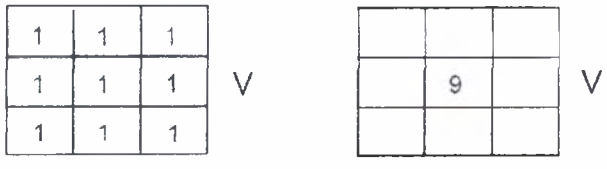


Figure 3-1: Call aggregation with a vehicle location.

The problem here is that timeliness measured on the aggregated system may greatly overestimate the timeliness of the actual system. Consider the 8-minute call coverage criteria typically used in Emergency Medical Service systems and assume that a single vehicle is located exactly 7.5 minutes from the center of the blocks directly to the right. In the aggregated problem, all calls would be considered covered since the vehicle is 7.5 minutes away. In the actual problem, calls in the left column and calls in blocks directly above and below the center block may not be covered as these travel values may be larger than 7.5 minutes. Similar examples using travel time or vehicle utilization as the criteria can easily be constructed. As the zones become smaller, the inaccuracies due to aggregation become smaller as well. Three specific types of errors may be defined:

- Errors in distance measurement for the call since the original call location is not the location of the aggregated calls.
- Errors in distance measurement due to not knowing the true location when a vehicle or facility is located at an aggregated zone.
- Errors in dispatching due to not knowing the correct distance from vehicles or bases to calls in aggregated zones.

As computing power increases and larger models can be formulated and solved, less aggregation is needed and this problem becomes less critical. At this time, aggregation can still cause problems in models that use coverage or travel time objectives. [17]

3.1.2 Demand Data Modeling and Prediction

A crucial issue in locating emergency response vehicles is data availability. Collection and analysis of the available data point out one of the main problems of the system. Models should be formulated in such a way that they use only data that can be collected, and they

must be robust, in the sense that the system designed by the model should not be too sensitive to small data errors. On the other hand, mathematical programming models almost always require less complete data to determine the necessary parameters. Furthermore, optimization models that are based on single or few objectives utilize a more simplified view of reality.

The ability to predict demand is of paramount importance. The typical approach is to tally past demand for each zone over some time period (a year or six months), and then assume that future demand will behave similarly to past demand. Similarly includes both quantity and spatial similarity. Even when the quantity of demand is changed, this is usually done in a proportional manner. [36]

3.1.3 Model Validity

Model validity refers to the model's ability to predict output and to make decisions that will work as well as predicted in the actual system. This is a key step in the modeling process. Unless the model makes valid predictions then the model will have little to no value.

Almost all models have "face validity" where the model looks reasonable to the casual observer. The next level is "replication validity." Here, the analyst inputs data on past operation of the actual system and the model replicates the operation of the system including:

- Predicting coverage and travel time close to those realized in the actual system.
- Making the same dispatching decisions as the actual system made.
- Predicting vehicle utilization close to that realized in the actual system.

The final level is "prediction validity" where the analyst inputs data for a future system and the model predicts how the future system will behave. Often future validity cannot be fully determined until the system is implemented. Hence if the model has face and

replication validity, then the decision maker is generally convinced of the quality of the model's output.

Chapter 4 Classification of Facility (Vehicles) Location Models

Facility location models can be classified according to their objectives, constraints, solutions, and other attributes.

Topological characteristics of the facility and demand sites lead to different location models including continuous location models, discrete network models, hub connection models etc. In each of these models, facilities can only be placed at the sites where it is allowed by topographic conditions.

The objective is an important criterion to classify the location models. Covering models aim to minimize the facility quantity while providing coverage to all demand nodes or maximize the coverage provided the facility quantity is pre-specified. Center models have an objective to minimize the maximum distance (or travel time) between the demand nodes and the facilities. P-median models attempt to minimize the sum of distance (or average distance) between the demand nodes and their nearest facilities.

Discrete location models assume that demands can be aggregated to a finite number of discrete points. Thus, we might represent a city by several hundred or even several thousand points or nodes (e.g., census tracts or even census blocks). Similarly, discrete location models assume that there is a finite set of candidate locations or nodes at which facilities can be sited. Continuous location models assume that demands are distributed continuously across a region much the way peanut butter might be spread on a piece of bread. These models do not

necessarily assume that demands are uniformly distributed, though this is a common assumption. Likewise, facilities can generally be located anywhere in the region in continuous location models.

Different solution methods result in different location models such as optimization models and descriptive models. Optimization models use mathematical approaches such as linear programming or integer programming to seek alternative solutions which trade off the most important objectives against one another. Descriptive models, in contrast, use simulation or other approaches to achieve successively enhanced location pattern until a solution with desired degree is achieved. Combined solution methods have also been developed by extending the descriptive models with optimization techniques to address dynamic and interactive location problems (e.g. mobile servers).

Features of facilities also divide location models into different kinds. For instance, facility restrictions can lead to models with or without service capacity and facility dependencies can result in models that take into account the facility cooperation or neglect it. These capacity limits can be for example the number of customers that can be attended by an ambulance system within a reasonable waiting time.

Location models can also be classified based on the demand patterns. If a model has elastic demand, then the demand in an area will vary (either increase or decrease) with different facility location decisions while a model with inelastic demand will not vary the demand pattern due to the facility location decisions.

Time horizon categorizes location models into static models and dynamic models. Static models optimize the system performance deciding all variables simultaneously. In

contrast, dynamic models consider different time periods with data variation across these periods, and give solutions for each time period adapting to the different conditions.

Another way to classify the location models is based on the features of the input parameters to the problems. In deterministic models, the parameters are forecast with specific values and thus the problems are simplified for easy and quick solutions. However, for most real-world problems, the input parameters are unknown and stochastic/probabilistic in nature. Stochastic/probabilistic location models capture the complexity inherent in real-world problems through probability distributions of random variables or considering a set of possible future scenarios for the uncertain parameters. [20]

Chapter 5 An Elaborate Description of the Basic Emergency Response Vehicles Location Models

The main section of this thesis concerns an elaborate description of the basic emergency response vehicles location models, mostly, in discrete space or networks, that are related to the public sector, such as ambulances, fire vehicles, police units.

5.1 Static and Deterministic Location Problems

These models assume that the nearest unit to a call for service is always available. The study of location theory formally began in 1909 when Alfred Weber considered how to position a single warehouse so as to minimize the total distance between it and several customers. Following this initial investigation, location theory was driven by a few applications which inspired researchers from a range of fields. Location theory gained renewed interest in 1964 with a publication by Hakimi , who sought to locate switching centers in a communications network and police stations in a highway system. To do so, Hakimi considered the more general problem of locating one or more facilities on a network so as to minimize the total distance between customers and their closest facility or to minimize the maximum such distance.

Since the mid-1960s, the study of location theory has flourished. The most basic facility location problem formulations can be characterized as both static and deterministic. These problems take constant, known quantities as inputs and derive a single solution to be

implemented at one point in time. The solution will be chosen according to one of many possible criteria (or objectives), as selected by the decision maker. [16, 30]

5.1.1 Covering Models

Covering models are based on the concept of acceptable proximity. In covering models, a maximum value is present for either distance or travel time. If a service is provided by a facility located within this maximum, then the service is considered acceptable. The service is equally good if provided by facilities at different distances, as long as both distances are smaller than this maximum value. Then, a customer is considered covered by the service, or just covered, if she/he has a facility sited within the preset distance or time.

Covering models can be classified according to several criteria. One of such criteria is the type of objective, which allows us to distinguish two types of formulations. In the first place, those seeking to minimize the number of facilities needed for full coverage of the population (Set Covering Models) and secondly, those that maximize covered population, given a limited number of facilities or servers (Maximum Covering Models).

5.1.1.1 Set Covering Model

The aim of this model is to locate a minimum number of servers needed to obtain mandatory coverage of all demands. In other words, each and every demand point has at least one server located within some distance or time standard r. The model positions the minimum possible number of emergency vehicles in such a way that the entire population has at least one of these vehicles initially located within the time or distance standard. Note that coverage is not affected by the fact that servers (vehicles) may be busy at times.

The formulation of the model is as follows:

$$\text{minimize} \quad \sum_{j \in J} x_j$$

$$\text{subject to:} \quad \sum_{j \in J_i} x_j \geq 1 \qquad \forall i \in I$$

$$x_j \in \{0,1\} \qquad \forall j \in J$$

where:

J = set of eligible facility sites

I = set of demand nodes

$$x_j = \begin{cases} 1 & \text{If a facility is located at node j} \\ 0 & \text{otherwise} \end{cases}$$

$J_i = \{j \in J : t_{ij} \leq r\}$

t_{ij} = shortest travel time from potential facility location j to demand node i

r = time standard for coverage

Note that J_i is the set of all those sites that are candidates for potential location of facilities, that are within time r of the demand node i. If a facility is located in any of them, demand node i becomes covered. The objective function minimizes the number of facilities required. Constraints state that the demand at each node i must be covered by at least one server located within the time or distance standard r. This model ignores several aspects of real-life problems, the most important probably being that once a vehicle is dispatched, some demand nodes are no longer covered. [4, 7, 8, 27, 39]

5.1.1.2 Maximal Covering Location Problem

The Maximal Covering Location Problem (MCLP) recognizes that mandatory coverage of all people in all occasions and no matter how far they live, could require excessive resources. Thus, MCLP does not force coverage of all demand but, instead, seeks

the location of a fixed number of facilities, most probably insufficient to cover all demand within the standards, in such a way that population or demand covered by the service is maximized. The formulation of the model is as follows:

$$\text{Maximize} \quad \sum_{i \in I} d_i y_i$$

$$\text{Subject to:} \quad \sum_{j \in J_i} x_j \geq y_i \qquad \forall i \in I \qquad (1)$$

$$\sum_{j \in J} x_j = p \qquad (2)$$

$$x_j \in \{0,1\} \qquad \forall j \in J \qquad (3)$$

$$y_i \in \{0,1\} \qquad \forall i \in I \qquad (4)$$

where additional:

d_i = demand at node i

p = the number of facilities to be deployed

$$y_i = \begin{cases} 1 & \text{If demand node i is covered} \\ 0 & \text{otherwise} \end{cases}$$

The objective function maximizes the number of covered demands. It is important to note that this model maximizes demands that are covered and not simply nodes. The first constraint states that demand node i cannot be counted as covered unless we locate at least one facility that is able to cover the demand node. The second constraint states that exactly p facilities are to be located and the other two constraints are standard integrality constraints. [4, 6, 7, 8, 27]

In most emergency systems, a fundamental issue is the amount of time a customer waits for service. This is the case of any public emergency services, either medical, fire fighting or police related. In the case of medical emergencies, there is a correlation between life loss risk and response time. Thus, it seems to be a good approach to assure medical attention of all calls within a time standard or, equivalently, have an available server within a standard distance of each and every customer. The same happens in the case of fire fighting services. Since it can be expected that loss of property increase with time, each type of company has to respond within its standard time. Many issues have to be considered in order to determine the performance of an emergency service. Response time is one of them. From the point of view of the geographical design of such a system, an important issue is the location of the depots, that is, the initial location of the emergency vehicles (servers). Another one is the number of servers and a third one is the availability of servers. Availability, is defined as the actual percentage of time the server is idle, as opposed to being on repair, or attending other calls.

Neither LSCM nor MCLP recognizes the fact that on occasions vehicles of several types may be dispatched to the scene of an incident. Also, even if only one vehicle type is used, solving MCLP alone may not provide a sufficiently robust location plan.

One of the first models developed to handle several vehicle types is the tandem equipment allocation model, or TEAM. It applies naturally to fire companies that operate with two types of equipment (pumpers and rescue ladders), but it is also relevant in an ambulance location context where Basic Life Support Units and Advanced Life Support Units are used. Denote by p^A and p^B the number of vehicles of types A and B available, let r^A and r^B be the coverage standards for each vehicle type, and define:

$J_i^A = \{j \in J : t_{ij} \leq r^A\}$

$J_i^B = \{j \in J : t_{ij} \leq r^B\}$

$x_j^A = \begin{cases} 1 & \text{If a vehicle of type A is located at node j} \\ 0 & \text{otherwise} \end{cases}$

$x_j^B = \begin{cases} 1 & \text{If a vehicle of type B is located at node j} \\ 0 & \text{otherwise} \end{cases}$

$y_i = \begin{cases} 1 & \text{If demand node i is covered by two types of vehicle} \\ 0 & \text{otherwise} \end{cases}$

The formulation of the model is as follows:

Maximize $\sum_{i \in I} d_i y_i$

Subject to:

$$\sum_{j \in J_i^A} x_j^A \geq y_i \qquad \forall i \in I$$

$$\sum_{j \in J_i^B} x_j^B \geq y_i \qquad \forall i \in I$$

$$\sum_{j \in J} x_j^A = p^A$$

$$\sum_{j \in J} x_j^B = p^B$$

$$x_j^A \leq x_j^B \qquad \forall j \in J$$

$$x_j^A , x_j^B \in \{0,1\} \qquad \forall j \in J$$

$$y_i \in \{0,1\} \qquad \forall i \in I$$

This model is a direct extension of MCLP except for constraints $x_j^A \leq x_j^B$ which impose a hierarchy between the two vehicle types. This constraint can of course be removed if

circumstances warrant it. In the facility-location, equipment-emplacement technique, or FLEET model constraints $x_j^A \leq x_j^B$ are relaxed, but only p location sites may be used. [37]

In any of the above models, coverage may become inadequate when vehicles become busy. A strategy employed in the case of a single vehicle type is to modify MCLP in order to provide better multiple coverage, without increasing the total number of vehicles beyond p. Two models with backup coverage, called BACOP1 and BACOP2, incorporate binary variables y_i equal to 1 if and only if demand point $i \in I$ is covered once by an ambulance within a coverage standard r, and binary variables u_i equal to 1 if and only if i is covered twice within r. The formulation of BACOP1 model is as follows:

Maximize $\sum_{i \in I} d_i u_i$

Subject to: $\sum_{j \in J_i} x_j \geq u_i + 1 \qquad \forall i \in I$

$\sum_{j \in J} x_j = p$

$0 \leq u_i \leq 1 \qquad \forall i \in I$

$x_j \geq 0 \qquad \forall j \in J$

The formulation of BACOP2 model is as follows:

Max $Z_1 = \sum_{i \in I} d_i y_i$

Max $Z_2 = \sum_{i \in I} d_i u_i$

Subject to: $\sum_{j \in J_i} \alpha_{ij} x_j \geq u_i + y_i \qquad \forall i \in I \qquad (1)$

$u_i - y_i \leq 0 \qquad \forall i \in I \qquad (2)$

$\sum_{j \in J} x_j = p \qquad (3)$

$$0 \le u_i \le 1 \qquad \forall i \in I$$

$$0 \le y_i \le 1 \qquad \forall i \in I$$

$$x_j \ge 0 \qquad \forall j \in J$$

J = set of eligible facility sites

I = set of demand nodes

d_i = demand at node i

p = the number of facilities to be deployed

$$y_i = \begin{cases} 1 & \text{If demand node i is covered at least once} \\ 0 & \text{otherwise} \end{cases}$$

$$u_i = \begin{cases} 1 & \text{If demand node i is covered at least twice} \\ 0 & \text{otherwise} \end{cases}$$

$$x_j = \begin{cases} 1 & \text{If a facility is located at node j} \\ 0 & \text{otherwise} \end{cases}$$

$$\alpha_{ij} = \begin{cases} 1 & \text{If } t_{ij} \le r \\ 0 & \text{otherwise} \end{cases}$$

t_{ij} = shortest travel time from potential facility location j to demand node i

r = time standard for coverage

The objective maximize first and second coverage. The first constraint says that coverage by a first and second server is not possible unless at least two servers are initially located in the neighbourhood. The second constraint reflects the fact that backup coverage can not be fulfilled without first coverage. The next constraint limits the number of servers to be deployed. [19]

5.1.2 Center Problems

An interesting feature of the LSCP, is that it can be used for solving the p-center problem, which consists on finding the locations of p facilities in such a way as to minimize the maximum distance between a customer and its allocated facility. There are several possible variations of the basic model. The "vertex" p-center problem restricts the set of candidate facility sites to the nodes of the network while the "absolute" p-center problem permits the facilities to be anywhere along the arcs. Both versions can be either weighted or unweighted. In the unweighted problem, all demand nodes are treated equally. In the weighted model, the distances between demand nodes and facilities are multiplied by a weight associated with the demand node. For example, this weight might represent a node's importance or, more commonly, the level of its demand. The vertex p-center problem can be formulated as follows:

Minimize W

Subject to:

$$\sum_{j \in J} x_j = p \qquad (1)$$

$$\sum_{j \in J} y_{ij} = 1 \qquad \forall i \in I \qquad (2)$$

$$y_{ij} - x_j \leq 0 \qquad \forall i \in I,\ \forall j \in J \qquad (3)$$

$$W - \sum_{j \in J} d_i s_{ij} y_{ij} \geq 0 \qquad \forall i \in I \qquad (4)$$

$$x_j \in \{0,1\} \qquad \forall j \in J \qquad (5)$$

$$y_{ij} \in \{0,1\} \qquad \forall i \in I,\ \forall j \in J \qquad (6)$$

where:

W = the maximum distance between a demand node and the facility to which it is assigned

d_i = demand at node i

s_{ij} = distance between demand node i and candidate site j

$$y_{ij} = \begin{cases} 1 & \text{if demand node i is assigned to a facility at node j} \\ 0 & \text{Otherwise} \end{cases}$$

The objective function minimizes the maximum demand-weighted distance between each demand node and its closest open facility. The first constraint stipulates that p facilities are to be located. The second constraint set requires that each demand node is assigned to exactly one facility. The third constraint set restricts demand node assignments only to open facilities. The forth constraint defines the lower bound on the maximum demand-weighted distance, which is being minimized. The fifth constraint set established the siting decision variable as binary. The sixth constraint set requires the demand at a node to be assigned to one facility only. This constraint can be replaced by $y_{ij} \geq 0 \ \forall i \in I, j \in J$ because the third constraint set guarantees that $y_{ij} \leq 1$. If some y_{ij} are fractional, we can simply assign node i to its closest open facility. This problem is adequate for its use in applications in the public sector, because it tends to generate certain equity in the access to facilities by their users. [16]

5.1.3 P-Median Problem

The p-Median Problem belongs to a class of formulations called minisum location models. The aim of this problem is to locate a fixed number of p facilities so as to minimize the weighted distance of the system. The p-Median problem can be formulated as follows:

Minimize $\quad \sum_{j \in J} \sum_{i \in I} d_i s_{ij} y_{ij}$

Subject to: $\quad \sum_{j \in J} y_{ij} = 1 \qquad \forall i \in I \qquad (1)$

$$y_{ij} - x_j \leq 0 \qquad \forall i \in I,\ \forall j \in J \qquad (2)$$

$$\sum_{j \in J} x_j = p \qquad (3)$$

$$x_j \in \{0,1\} \qquad \forall j \in J$$

$$y_{ij} \in \{0,1\} \qquad \forall i \in I,\ \forall j \in J$$

Where:

d_i = demand at node i

s_{ij} = distance between demand node i and candidate site j

$$y_{ij} = \begin{cases} 1 & \text{if demand node i is assigned to a facility at node j} \\ 0 & \text{Otherwise} \end{cases}$$

$$x_j = \begin{cases} 1 & \text{If a facility is located at node j} \\ 0 & \text{otherwise} \end{cases}$$

The objective function minimizes the demand weighted total distance. This is equivalent to minimizing the demand weighted average distance since the total demand is a constant. The first constraint states that each demand node must be assigned to exactly one facility site. The second constraint stipulates that demand nodes can only be assigned to open facility sites. The third constraint states that we are going to locate exactly p facilities. The last two constraints are standard integrality constrains.

5.2 Dynamic Location Problem

The models previously described tend to be "single use" models. A user would solve the model for a single data set of demands, travel times, and service times, and obtain insight on good sets of locations for that data set. This is problematic in that the data is typically not stationary and has dramatic changes over the day, the week, and even the year. One approach for dealing with the dynamic nature of the problem is to break the week into 168 hourly

periods and solve the model for each hour. Here, a user will have to integrate solutions so that the system runs smoothly and is not jumpy with vehicles changing locations repeatedly. Also, one can use the models and do pre-planning for atypical situations. For example, if 25% of the vehicles are busy, one could solve a model with 25% less capacity and see how the system should be designed. This solution can now be used to help in deciding how to re-deploy. So, the typical strategy for dealing with the dynamic nature is to use the static models and do a great deal of experimentation to pre-plan for contingent situations. Unfortunately, one cannot anticipate every possible situation and one must still figure out how to integrate and implement the solutions from the different model runs.

When siting emergency vehicles, relocation decisions must periodically be made in order not to leave areas unprotected. This was recognized by Kolesar and Walker in 1974 [21] who designed a relocation system for fire companies. With the development of faster heuristics and advanced computer technologies, it is now possible to quickly solve an ambulance location problem in real-time. This means that a new ambulance redeployment strategy can be recomputed at any time t, using the available information.

One such model exists in the area of ambulance relocation and it was developed as follows. In addition to the standard coverage and site capacity constraints, the model takes into account a number of practical considerations inherent to the dynamic nature of the problem: a) vehicles moved in successive redeployments cannot always be the same, b) repeated round trips between the same two location sites must be avoided, c) long trips between the initial and final location sites must be avoided.

The ambulance relocation problem is solved at each instant t at which a call is registered. The dynamic aspect of the redeployment model is captured by time dependent

constants $M^t_{j\ell}$ equal to the cost of repositioning, at time t, ambulance ℓ from its current site to site $j \in J$. This includes the case where site j coincides with the current location of the ambulance, i.e., $M^t_{j\ell} = 0$. The constant $M^t_{j\ell}$ captures some of the history of ambulance ℓ. If it has been moved frequently prior to time t, then $M^t_{j\ell}$ will be larger. If moving ambulance ℓ to site j violates any of the above constraints, then the move is simply disallowed. Binary variables $y_{j\ell}$ are equal to 1 if and only if ambulance ℓ is moved to site j. The authors defined γ_{ij} as a binary coefficient indicating whether $t_{ij} \leq r_1$ ($\gamma_{ij} = 1$) or not ($\gamma_{ij} = 0$), δ_{ij} as a binary coefficient indicating whether $t_{ij} \leq r_2$ ($\delta_{ij} = 1$) or not ($\delta_{ij} = 0$) and x_i^k as a binary variable equal to 1 if and only if demand node i is covered at least k times. The model at time t can be formulated as follows:

Maximize $$\sum_{i=1}^{n} d_i x_i^2 - \sum_{j=1}^{m}\sum_{\ell=1}^{p} M^t_{j\ell} y_{j\ell}$$

Subject to:

$$\sum_{j=1}^{m}\sum_{\ell=1}^{p} \delta_{ij} y_{j\ell} \geq 1 \qquad \forall i \in I \qquad (1)$$

$$\sum_{i=1}^{n} d_i x_i^1 \geq \alpha \sum_{i=1}^{n} d_i \qquad (2)$$

$$\sum_{j=1}^{m}\sum_{\ell=1}^{p} \gamma_{ij} y_{j\ell} \geq x_i^1 + x_i^2 \qquad \forall i \in I \qquad (3)$$

$$x_i^2 \leq x_i^1 \qquad \forall i \in I \qquad (4)$$

$$\sum_{j=1}^{m} y_{j\ell} = 1 \qquad (\ell = 1,\ldots\ldots p) \qquad (5)$$

$$\sum_{\ell=1}^{p} y_{j\ell} \leq p_j \qquad \forall j \in J \qquad (6)$$

$$x_i^1, x_i^2 \in \{0,1\} \qquad \forall i \in I$$

$$y_{j\ell} \in \{0,1\} \qquad (j \in J,\ \ell = 1,\ldots\ldots p)$$

Apart from the variables y_{jt}, all variables, parameters and constraints of this model can be interpreted as in the static case. The objective function is to maximize the backup coverage demand i.e., the proportion of the demand covered by at least two vehicles within a radius r_1, minus a relocation cost. The first term of the objective function is particularly appropriate in a real time context since the zone covered by a vehicle assigned to a call may still be covered by another ambulance after the call has been serviced. The second term ensures that the location plan remains fairly stable throughout the day. The model is truly dynamic since it incorporates new information on the state of the system received at each period t. In this model the first and second constraints ensure the single and the double coverage requirements. The absolute covering first constraint states that all demand must be covered within r_2 units. The second and third constraints express the relative covering requirements. The second constraint imposes that a proportion α of all demand is covered whereas the third constraint states that the number of ambulances located within r_1 units should be at least one if $x_i^1 = 1$ or at least two if $x_i^2 = x_i^1 = 1$. The forth constraint ensures that a demand point cannot be covered twice if it is not covered at least once. The fifth constraint specifies that each available ambulance must be assign to a potential location site. The sixth constraint defines an upper bound on the number of vehicles waiting at a location site. [14]

5.3 Probabilistic Location Problems

The dynamic models described in the previous section attempt to locate facilities over a specified time horizon in an optimal or near-optimal manner. While capturing more of the complexity inherent in real world problem instances than static and deterministic formulations, these models assume that input parameters are known values or that they vary deterministically over time.

Probabilistic location problems deal with the stochastic nature of real-world systems. In these systems, models capture the stochastic aspects of facility location through explicit consideration of the probability distributions associated with modeled random quantities. Parameters, such as for example travel times, the location of clients, demand and the availability of servers are treated as random variables. The objective is to determine robust server/facility locations that optimize a given utility function, for a range of values of the parameters under consideration.

Researchers incorporate these distributions into standard mathematical programs, while others use them within a queueing framework.

5.3.1 Functions of random variables

Given an experiment with a sample space and a probability assignment over the sample space, a random variable is a function that assigns a numerical value to each finest-grained outcome in the sample space.

Each probabilistic modelling experiment can be approached using the following four steps:

- Define the random variables of interest
- Identify the joint sample space
- Determine the joint probability distribution over the sample space
- Work within the sample space to determine the answers to any questions about the experiment

When examining a system we know by hypothesis or measurement the probability law of one or more random variables and we wish to obtain the probability laws of other random variables that can be expressed in terms of the original random variables. The random

variables in the second set are functions of the random variables in the first set. This is a problem of derived distributions, since we must derive the joint probability distribution for the random variables in the second set. Derived distribution problems can arise with discrete, continuous, or mixed random variables.

One technique for deriving distributions, is the "never-fail" method. Virtually all of the work associated with this method occurs in the joint sample space of the original random variables. Suppose that the original set of random variables is given by $\{X_1, X_2, ..., X_N\}$ with joint cdf $F_{X1, X2,...,XN}(\cdot)$. Suppose that there are M random variables $Y_1, Y_2, ..., Y_M$, each of which can be expressed as a function of $X_1, X_2, ..., X_N$, namely $Y_i = g_i(X_1, X_2, ..., X_N)$, i = 1, 2, ..., M. Then the never-fail method, called the cumulative distribution method, allows computation of the joint cumulative distribution function for the Y_i's, as follows:

$$F_{Y_1, Y_2, \ldots Y_M}(y_1, y_2, \ldots y_M) \equiv P\{Y_1 \le y_1, Y_2 \le y_2, \ldots Y_M \le y_M\}$$

a. Identify the set of points in the original $(X_1, X_2, ..., X_N)$ sample space that corresponds to the joint event

$$\{Y_1 = g_1(X_1, X_2, \ldots X_N) \le y_1, Y_2 = g_2(X_1, X_2, \ldots X_N) \le y_2, \ldots, Y_M = g_M(X_1, X_2, \ldots X_N) \le y_M\}$$

b. For each set of values for the y's, $[y_1, y_2, \ldots, y_M]$, determine by summation or integration the probability in the $(X_1, X_2, \ldots X_M,)$ sample space of this joint event, thereby obtaining $F_{Y1,Y2,\ldots Ym}(y_1, y_2, \ldots y_M)$ where $-\infty < y_1, y_2, \ldots y_M < +\infty$

If the random variables are continuous, we can find the joint pdf for $\{Y_1, Y_2, \ldots Y_M\}$ by taking partial derivatives of $F_{Y1,Y2,\ldots YM}(\cdot)$ with respect to each of its arguments,

$$F_{Y_1, Y_2, \ldots Y_M}(y_1, y_2, \ldots y_M) = \frac{\partial^M}{\partial y_1 \partial y_2 \ldots \partial y_M} F_{Y_1, Y_2, \ldots Y_M}(y_1, y_2, \ldots, y_M)$$

If they are discrete, the pmf is found simply by using the cdf and subtracting appropriate successive values.

5.3.2 Expected travel distance for a public safety vehicle to reach a random emergency incident in a straight highway of unit length

As we mentioned above, the time it takes for municipal emergency vehicles, such as fire engines, police cars, and ambulances, to respond to calls for service is an important and widely used indicator of the performance of emergency service systems. Most municipalities, know very little about how quickly their emergency units respond and how travel times and travel speeds vary with response distance, time of day, and region of the city.

One approach to determine the number of units to locate in a region is to estimate the average travel time as a function of the number of units and find the number of units needed to achieve a target average travel time.

One of the most important indicators of the performance of any emergency service system is response time, the time interval between the receipt of a call for service and the arrival of an emergency unit at the scene of the incident. Since response time can have a significant impact on the loss of life and property at an emergency, it is used as a principal measure of effectiveness in many models developed for analyzing the deployment of emergency vehicles.

Response time can be divided into three components: a) dispatch time, b) turnout time and c) travel time. Dispatch time is the time elapsed between the receipt of a call for service and the dispatch of the service unit. Turnout time, which is a factor only if the emergency unit is not immediately ready to respond when dispatched, is the time elapsed between the dispatch of a unit to a call for service and the departure of the emergency unit for the scene.

This component is relatively constant, since departure preparations generally consume the same amount of time regardless of the type of call. The third component of response time is travel time, or the time it takes for the emergency unit to arrive at the incident after it begins its response. [10]

Suppose that a public safety vehicle travels ***back and forth*** along a straight highway, the travelling perhaps to find motorists in need of assistance. Also, along this highway accidents can occur that create a need for on-scene assistance by the vehicle. The vehicle is dispatched by radio to these accidents. We are interested in determining the probability law of the travel distance for the public safety vehicle to reach a random emergency incident.

It requires that we do four things to model the experiment:

- Define the random variables of interest
- Identify the joint sample space
- Determine the joint probability distribution over the sample space
- Work within the sample space to determine the answers to any questions about the experiment

1. *Random variables.* Suppose that the highway is of unit length. Then the two key random variables would be:

X_1 = location of the emergency incident, $0 \leq X_1 \leq 1$

X_2 = location of the safety vehicle at the moment of dispatch, $0 \leq X_2 \leq 1$

The travel distance D can be expressed as a function of X_1 and X_2, $D = |X_1 - X_2|$

2. *Joint sample space.* The joint sample space is the unit square in the positive quadrant ($0 \leq X_1 \leq 1$, $0 \leq X_2 \leq 1$).

3. *Joint probability distribution.* We assume that the locations of the public safety vehicle and the emergency incident are uniformly, independently distributed over the highway. Naturally, the analysis could also proceed with an alternative set of assumptions. Since we are now dealing with strictly continuous random variables, we will work with the joint probability density function, which is:

$$f_{x_1,x_2}(x_1,x_2) = f_{x_1}(x_1)f_{x_2}(x_2) = \begin{cases} 1 \cdot 1 = 1 & 0 \leq X_1, X_2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

4. *Work in the sample space.* This is the point at which the never-fail method for deriving distributions comes into play. We want the probability law of:

$$D = |X_1 - X_2| = \text{"travel distance"}$$

To apply the never-fail method for finding the cdf of D, $F_D(y)$, we first locate the region in the (X_1, X_2) sample space corresponding to the event $(D \leq y)$. Formally, the steps are written as follows:

$$F_D(y) \equiv P\{D \leq y\} = P\{|X_1 - X_2| \leq y\}$$

To remove the absolute value operator, we consider two cases separately:

case 1: $X_1 \geq X_2$ and case 2: $X_1 < X_2$.

For the first case, $D = X_1 - X_2$ and experimental values x_1 and x_2 of X_1 and X_2, respectively, must lie between the line $x_2 = x_1$ and $x_2 = x_1 - y$. For the second case, $D = X_2 - X_1$, and experimental values of X_1 and X_2 must lie between the line $x_2 = x_1$ and $x_2 = x_1 + y$. Consideration of these two cases gives rise to the shaded region in the sample space in Figure

5-1. Once we have determined such a region, we have identified the set of points corresponding to the event of interest D ≤ y, thereby completing step a of the never-fail method.

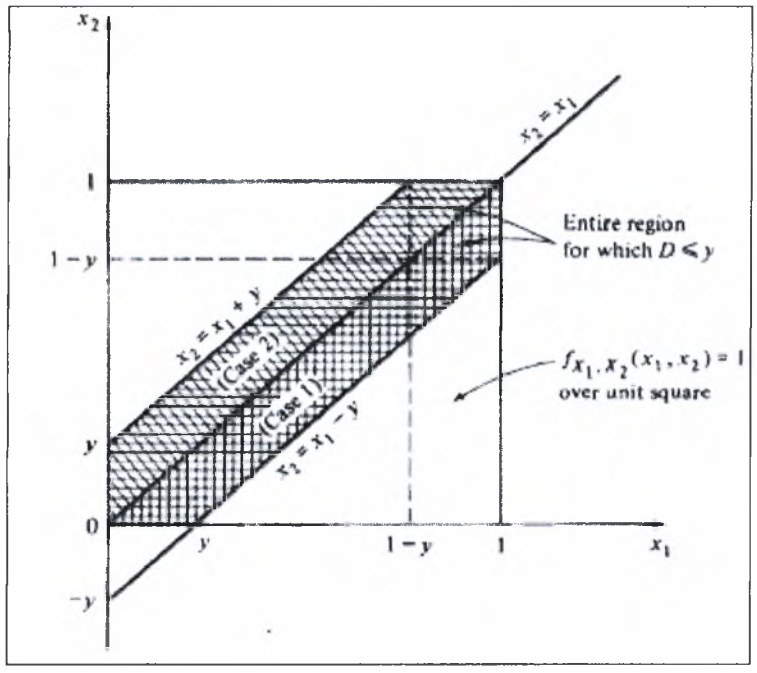


Figure 5-1: Joint (X_1, X_2) sample space, randomly positioned incident and vehicle.

Step b of the never-fail method requires that we integrate $f_{x1,x2},(\cdot)$ over the set of points in the shaded region to obtain $F_D(y)$. Since the joint X_1, X_2 pdf is uniform over the unit square, we can perform the integration by computing areas in the sample space. (Conceptually, each area is multiplied by "l," the height of the pdf at that point, to yield a probability measured as a volume.) By computing areas of the triangles not in the shaded region,

$$F_D(y) = 1 - 2(1/2)(1-y)^2 \qquad 0 \leq y \leq 1$$

we have now completed step b of the never-fail method. Should we desire the pdf of D, we differentiate, obtaining:

$$f_D(y) = \frac{dF_D(y)}{dy} = \begin{cases} 2(1-y) & 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

From the pdf (or cdf) we can determine anything that is desired concerning D. So, the expected value (or mean value) of D is:

$$E[D] = \int_{-\infty}^{+\infty} y f_D(y) dy = \int_0^1 y2(1-y) dy = \frac{1}{3}$$

and the variance of D is:

$$\sigma_D^2 = E[(D - E[D])^2] = E[D^2] - (E[D])^2 = \frac{1}{18}$$

Suppose that a system administrator is interested in knowing the effects on travel distance of ***prepositioning*** the public safety vehicle at the center of the interval depicting the highway, thus fixing X_2. Then the joint sample space is the straight line indicated in Figure 5-2. If the new travel distance is $D'=|X_1-1/2|$, the region for which $(D' \le y)$ is the line segment of length 2y centered at $X_1=1/2$. Integrating the uniform pdf of X_1, we have:

$$F_{D'}(y) = P\{D' \le y\} = P\{ | X_1 - 1/2 | \le y\} = 2y \ (O \le y \le 1/2).$$

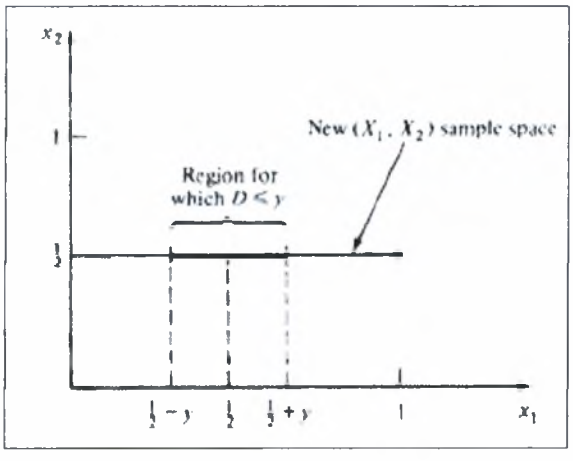


Figure 5-2: Joint (X_1, X_2) sample space, randomly positioned incident, fixed position vehicle.

Thus, the pdf of D' is: $$f_{D'}(y) = \begin{cases} 2 & 0 \le y \le 1/2 \\ 0 & \text{otherwise} \end{cases}$$

The mean and variance are: $E[D'] = \frac{1}{4}$ and $\sigma^2_{D'} = \frac{1}{48}$

Thus, a change in deployment policy resulting in an vehicle prepositioned at the center of its service area rather than randomly patrolling its service area reduces mean travel distance by 25 percent, the variance of the travel distance by 62.5 percent, and, perhaps important in "worst-case" analyses, the maximum possible travel distance by 50 percent.

5.3.3 A general case of expected travel distance in two dimension area

Travel distance is much easier to estimate than is travel time. As a result, many emergency service standards are based on distance. There are many simple methods that can be used to estimate distance (D) for a specific response.

For example, the distance can be measured on a map by following the actual route of response. An easy method to estimate distances by computer involves superimposing a rectangular grid on a map of the city and storing this grid in the computer. Then, any point in the city, such as a street intersection, firehouse, or fire alarm box, can be identified by a pair of grid coordinates (x, y). The distance between two points specified by (x_1, y_1) and (x_2, y_2) can then be estimated using a function of these coordinates.

So, the Euclidean distance (the straight line distance between the two points) is given by $D_E = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$, while the right angle distance (the distance calculated as if all streets in the city intersected at right angles) is given by $D_R = |x_2 - x_1| + |y_2 - y_1|$.

Let (X, Y) and (X_1, Y_1) indicate, respectively, the location of calls for service and of the response unit in a district R of area A. Denote by $f_{x,y,x1,y1}(x,y,x_1,y_1)$ the joint pdf for random

variables X, Y, X_1 and Y_1 and by $D = d[(X_1, Y_1), (X, Y)]$ the mathematical relationship for the distance between (X_1, Y_1) and (X, Y) [for example, $D = \sqrt{(X_1 - X)^2 + (Y_1 - Y)^2}$, for Euclidean distances]. Then, the expected travel distance in the district is:

$$E[D] = \iiiint_{\text{over R}} d[(x_1, y_1), (x, y)] f_{X,Y,X_1,Y_1}(x, y, x_1, y_1)\, dx_1\, dy_1\, dx\, dy \qquad (1)$$

Note that the joint pdf for the coordinates of the incident and of the service unit can be made to reflect not only nonuniformities in the distribution over R but also possible dependencies between the locations of incidents and of the service unit.

The former expression can be extended to the case where N response units are located in district R. If (X_i, Y_i) indicates the location of the i^{th} response unit (i = 1, 2,.. N) and (X, Y) the location of an incident, then the distance between the incident and the closest response unit can be written:

$$D_N = \text{Min}\ \{d[(X_1, Y_1), (X, Y)], \ldots, d[(X_N, Y_N), (X, Y)]\}$$

Since D_N, is then a function of the random variables X, Y, X_1, Y_1, X_2,Y_2,. . . . , X_N, Y_N:

$$E[D_N] = \int_{\text{over R}} \cdots \int \text{Min}\{d[(x_1, y_1), (x, y)], \ldots, d[(x_N, y_N), (x, y)]\} \bullet \qquad (2)$$

$$f_{X,Y,X_1,Y_1,\ldots,X_N,Y_N}(x, y, x_1, y_1, \ldots, x_N, y_N)\, dx\, dy \ldots dy_N$$

where $f_{x, y, x1, y1, \ldots, xn, yn}(x, y, x_1, \ldots, y_n)$ is obviously the joint pdf for the coordinates of the incident and the N response units. Thus, in both (1) and (2) we have expressed expected travel distance as the expected value of a function of random variables whose joint pdf is known. The problem of computing the expected travel distance in the general case is, therefore, no more (or less) difficult than working with any other function of these random variables.

Obviously, in practice, there are severe limitations on how far one can go in deriving such exact expressions for E[D]. Problems become mathematically intractable as the number of random variables increases or as the shape of R and/or the joint pdf for the random variables becomes more complex. In many cases, however, all is not lost as long as one is willing to settle for good approximations rather than exact results. This is true any time the response units are stationary at known locations, no matter what the number, N, of these units is (and for practically any pdf for the spatial distribution of incidents/demands as well as for any shape of the district of interest). It is also true, for any value of N, in the case of mobile response units as long as this approach can also be generalized to expected distances to other than the closest unit (e.g., to the k^{th} closest unit). Sub-districts of responsibility have been defined in such a way that each sub-district of R is served exclusively by a very small number of mobile units. In such instances, the following three-step approach will work:

STEP 1: Divide the district.R into several (possibly many) nonoverlapping parts, which we shall call "zones." Each zone must have the following two properties:

a. Its shape must be approximately rectangular, triangular, circular, or any other easy-to-work-with configuration.

b. The pdf for the spatial distribution of incidents/demands within each zone must be approximately uniform (or that pdf can be approximated by some other sufficiently simple expression as to permit easy mathematical manipulation).

STEP 2: We compute all intrazone and zone-to-response unit expected distances. STEP 3: Multiply the expected distances computed in Step 2 by appropriate probabilities to obtain overall expected travel distances for district R.

Note that each zone in Step 1 can have an individual shape with its "own" pdf for the distribution of incidents. Note also that the greater the degree of accuracy desired, the larger the number of district zones should be (to approximate better the shape of the district R and the pdf for the spatial distribution of incidents). In fact, the three- step approach outlined above is very similar to the approach that a computer would follow in order to compute numerically the integrals in expressions (1) and (2).

So far, we focused on case where incidents are not uniformly distributed and the district itself does not have a nice rectangular (or circular, triangular, etc.) shape.

In the following paragraphs we estimate expected travel distances and times to and from incidents in districts with relatively regular ("fairly compact and fairly convex") geometries and uniform distribution of incidents over the districts.

To find the district dimensions which lead to the minimum expected travel distance, we must keep for example, the area of the response district $A_o = X_o Y_o$, Figure 5-3, constant and $E[D] = \frac{1}{3}(X_o + Y_o)$ is minimized subject to the condition $Y_o = A_o/X_o$. Without this constant, a zero area (point) district would be optimal, an obviously infeasible result considering that the collection of districts in a city must usually cover the entire city (which has fixed positive area). Not surprisingly, E[D] is minimized when the rectangle becomes a square. In that case, we have $X_o = Y_o = \sqrt{A_o}$ and $E[D] = \frac{2}{3}\sqrt{A_o}$.

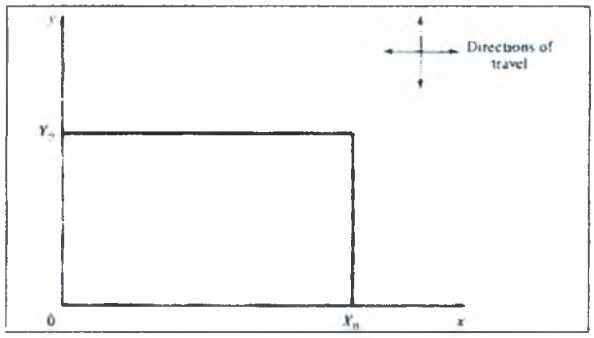


Figure 5-3: Rectangular Response Area.

More generally, if the effective travel speeds in the x-direction and the y-direction, v_x and v_y, are independent of travel distance, the expected travel time, $E[T] = \frac{1}{3}\left(\frac{X_o}{v_x} + \frac{Y_o}{v_y}\right)$ is minimized when $\frac{Y_o}{v_y} = \frac{X_o}{v_x} = \sqrt{\frac{A_o}{v_x v_y}}$ in which case $E[T] = \frac{2}{3}\sqrt{\frac{A_o}{v_x v_y}}$. The optimal shape of the district, is the one for which it takes as much time to traverse the district from "east to west" as from "north to south."

The expressions for E[D] and E[T] turn out to be "robust" (i.e., rather insensitive to the exact values of X_o and Y_o). We suppose that $X_o = \alpha Y_o$ where α is a positive constant. We assume $\alpha > 1$ and, we set $A_o = X_o Y_o$. Then E[D] can be written as:

$$E[D] = \frac{(\alpha+1)\sqrt{A_o}}{3\sqrt{\alpha}} = \frac{2}{3}\sqrt{A_o} + \frac{(\sqrt{\alpha}-1)^2}{3\sqrt{\alpha}}\sqrt{A_o}$$

The second term $\frac{(\sqrt{\alpha}-1)^2}{3\sqrt{\alpha}}\sqrt{A_o}$ is the amount by which E[D] deviates from its minimum value $E[D] = \frac{2}{3}\sqrt{A_o}$. For $\alpha = 1{,}5$ that term becomes equal to $0.014\sqrt{A_o}$, (i.e., E[D] is only about 2 percent greater than its minimum value). Even for $\alpha = 4$, E[D] is only 25 percent more than its minimum value.

Results such as those of $E[D] = \frac{2}{3}\sqrt{A_o}$ and $\frac{Y_o}{v_y} = \frac{X_o}{v_x} = \sqrt{\frac{A_o}{v_x v_y}}$ can be derived for various district shapes. The first three columns of Table 5-1 summarize the equivalents of $E[D] = \frac{2}{3}\sqrt{A_o}$, for a square district, a square district rotated by 45' with respect to the right-angle directions of travel, and a circular district. The following four cases are included:

1. Euclidean (straight-line) travel when the response unit is randomly and uniformly positioned in the district.
2. Case with right-angle travel.
3. Euclidean travel with the response unit located at the center of the district.
4. Case with right-angle travel.

In all cases it is assumed that the locations of requests for service are uniformly distributed in the district and independent of the location of the service unit. When the constants in Table 5-1 are multiplied by $\sqrt{A_o}$, the square root of the area of the district in question, E[D] is obtained. In some instances (e.g., a square district with a randomly positioned response unit and Euclidean travel) the constant of interest is not known exactly and the best known approximation, to two-decimal-place accuracy, is shown.

The three district geometries included in Table 5-1 are "special cases" of rectangular, diamond-shaped, and elliptic districts. If one varies the district dimensions of each type while constraining district area to equal a constant A_o, E[D] is minimized by the symmetric geometries represented in Table 5-1.

	Metric in use	Shape of district			
		Square	Perfect, four sided diamond	Circle	Approximation for "fairly compact and fairly convex" areas
Response unit is randomly positioned in the district	Euclidean travel	0,52	0,52	$\frac{128}{45\pi\sqrt{\pi}} = 0{,}511$	0,52
	Right angle travel	$2/3 = 0{,}667$	$\frac{14\sqrt{2}}{30} = 0{,}660$	$\frac{4 \cdot 128}{\pi \cdot 45\pi\sqrt{\pi}} = 0{,}650$	0,67
Response unit is located at the center of the district	Euclidean travel	$\frac{\sqrt{2} + \ln(1+\sqrt{2})}{6} = 0{,}383$	$\frac{\sqrt{2} + \ln(1+\sqrt{2})}{6} = 0{,}383$	$\frac{2}{3\sqrt{\pi}} 0{,}376$	0,38
	Right angle travel	$1/2 = 0{,}5$	$\sqrt{2}/3 = 0{,}471$	$\frac{4 \cdot 2}{\pi \cdot 3\sqrt{\pi}} = 0{,}479$	0,50

Table 5-1: Proportionality constants for determining mean travel distances.

It can be seen from Table 5-1 that, for any given district area A, E[D] is very insensitive to the exact geometry of the district. This can be confirmed by deriving E[D] for other possible district geometries, such as equilateral triangles or piece-of-pie-like sectors of circles. Moreover, for any given district geometry, the value of E[D] is insensitive to changes of the dimensions of the district that might make it appear to deviate appreciably from its optimum shape.

From these observations it can be concluded that we can use the first three columns of Table 5-1 to infer similar approximate expressions for E[D] that apply to districts of any shape as long as (1) one of the dimensions (e.g., "length") is not much greater than the other dimension (e.g., width), and (2) major barriers or boundary indentations do not exist in the district. Districts that satisfy both of the conditions above will be called here, informally, "fairly compact and fairly convex districts." We can now state the following: For fairly compact and fairly convex districts and for independently and spatially uniformly distributed requests for service, $E[D] = c \cdot \sqrt{A_o}$ where A_o, is the area of the district and c is a constant

that depends only on the metric in use and on the assumption regarding the location of the response unit in the district.

The last column of Table 5-1 lists values that can be used for c in $E[D] = c \cdot \sqrt{A_o}$ for the four combinations of response unit locations and metrics. In all cases, we have selected the largest value of c listed in each row of the three leftmost columns of Table 5-1.

When the effective travel speed is independent of the distance covered, one can use the constants in the fourth column of Table 5-1 to approximate the expected travel time, E[T], as well. In that case we have $E[T] = \frac{c}{v}\sqrt{A_o}$ in the case of Euclidean travel (assuming that the effective travel speed v is independent of the direction of travel) and $E[T] = c\sqrt{\frac{A_o}{v_x v_y}}$ for right-angle travel. In this latter case, the district "compactness" statement requires that E[Teast-west] ≈E[Tnorth-south]. That is, it takes on the average about as much time to traverse the district from east to west as from north to south.

In 1974 Kolesar and Blum [22] have shown that E[D] in a region is inversely proportional to the square root of the number of units per unit area. More formally, if the coordinates of each point (x, y) in the district of interest are multiplied by $\sqrt{m}$ (m > 1) [i.e., point (x, y) now becomes point ($\sqrt{m}$ x, $\sqrt{m}$ y)], then the area of the district increases m-fold but the length, L, of any given route between the pair of points (x_1, y_1,) and (x_2,, y_2,) in the original district-becomes equal to $\sqrt{m}$ L in the expanded district.

Equivalently, we can state that E[D] and E[T] must be proportional to the inverse of the square root of the density of response units in a district, for districts with more than one response unit. That is, if a district of area A is divided into n approximately equal fairly

convex and fairly compact subdistricts of responsibility (whose shapes may vary), then E[D] $= c\sqrt{\frac{A_o}{N}} = \frac{c}{\sqrt{y}}$ where y denotes the spatial density of service units and N denotes the number of available units.

In most practical situations, the effective travel speed of urban response units depends on travel distance: longer trips, in general, are taken at a higher average speed than are shorter trips. It is therefore desirable to develop expressions for E[T] that take into consideration some types of functional relationships between travel time and travel distance [unlike expressions $E[D] = \frac{2}{3}\sqrt{A_o}$ and $E[T] = c\sqrt{\frac{A_o}{v_x v_y}}$, which assumed that effective travel speed remains constant with distance]. One plausible model is the following. Let us assume that urban service vehicles responding to a call, first go through an acceleration stage (perhaps while maneuvering their way through side streets, turns, etc.) until they reach a cruising speed that they maintain through the middle stage of the trip (while, perhaps, traveling on highways, thoroughfares, etc.) up to the final stage of it, during which they decelerate to a stop. Let us further assume that during the initial and final stages, vehicles accelerate (or decelerate) at a constant rate of a miles/min and that during the middle stage, travel is at a constant cruising speed of v, miles/min.

For trips of length less than $2d_c$, (where $d_c = u^2_c/2\alpha$ is the distance needed to reach cruising speed) the cruising speed will neverbe reached; this is not the case when the travel distance D is greater than $2d_c$. Using the well-known physical relationships for accelerated and constant speed travel ($D = \alpha t^2/2$ and $D = ut$), it is then easy to conclude that the conditional expected travel time $E[T \mid D = d]$ for any given travel distance is:

$$E[T|D = d] = 2\sqrt{\frac{d}{\alpha}} \qquad \text{for } d \le 2d_c$$

$$E[T|D = d] = \frac{d - 2d_c}{u_c} + \frac{2u_c}{\alpha} = \frac{d}{u_c} + \frac{u_c}{\alpha} \qquad \text{for } d > 2d_c$$

One can obviously think of many other physical scenarios that would lead to different expressions for E[T | D = d]. A considerable amount of field data, however, suggests that these two relationships often provide truly excellent approximations for many urban services- see, for instance.

An expression for the unconditional expected travel time, E[T], can now be written:

$$E[T] = \int_0^\infty E[T|D = x] f_D(x)\, dx = \int_0^{2d_c} 2\sqrt{\frac{x}{\alpha}}\, f_D(x)\, dx + \int_{2d_c}^\infty \left(\frac{u_c}{\alpha} + \frac{x}{u_c} \right) f_D(x) dx$$

5.3.4 Coverage

In deployment applications one may be less interested in the expected value of some quantity, say travel time, than in the fraction of the city which receives "adequate coverage" by the service. Coverage is usually defined in terms of an inequality, such as travel time being less than or equal to i.e. 4.0 minutes.

Let us define coverage over a convenient interval, say [0, 1] and X ≡ set of points in [0,1] which are covered and μ(X) ≡ 'length' of the set X. If the covered points are as in Figure 5-4, then:

$$\mu(X) = \frac{1}{3} + \left(\frac{2}{3} - \frac{1}{2} \right) + \left(1 - \frac{3}{4} \right) = \frac{3}{4}$$

The points in X are usually determined according to some probabilistic process, and we wish to compute the expected value of μ(X).

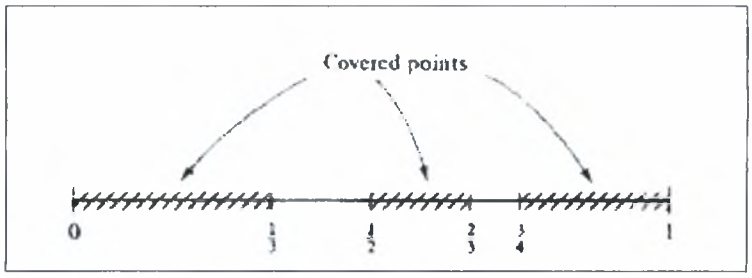


Figure 5-4: Set of points in [0,1] which are covered.

For example, suppose that we have N emergency vehicles distributed independently and uniformly over the interval [- δa, 1 + δa], the extensions beyond [0, 1] being added to avoid boundary problems. Emergency incidents are distributed uniformly on [0, 1] and are independent of vehicles positions. We want to know the expected amount of the interval [0, 1] which is "covered" by emergency vehicles, where a point is said to be covered if at least one vehicle is within a distance a/2 of the point. Suppose that the probability that any point x on [0,1] is covered by a particular emergency vehicle is $p_1(x) \equiv p_1$. Then, the probability that any point x on [0,1] is covered by at least one vehicle is:

$$p_N(x) = 1 - P\{x \text{ not covered by any vehicle}\} = 1 - (1 - p_1)^N \equiv p_N$$

We define a set indicator random variable as:

$$S(x) \equiv \begin{cases} 1 & \text{if } x \text{ is covered} \\ 0 & \text{otherwise} \end{cases}$$

and divide [0,1] into I intervals, where interval I has length $\Delta x = 1/I$. Then:

$$\mu(X) \cong \sum_{i=1}^{I} S(i\Delta x)\Delta x$$

and taking expected values:

$$E[\mu(X)] \cong E[\sum_{i=1}^{I} S(i\Delta x)\Delta x] = \sum_{i=1}^{I} E[S(i\Delta x)]\Delta x = \sum_{i=1}^{I} [0 \cdot (1 - p_N) + 1 \cdot p_N]\Delta x = p_N \cdot 1$$

Hence,

$$E[\mu(X)] = 1 - (1 - p)^N$$

The approximate summation above becomes an integral when $I \to \infty$, $\Delta_x \to 0$. Note that the solution behaves as we expect, namely diminishing marginal returns (in terms of extra expected area covered) with each additional vehicle. Generalizing the foregoing argument, if p(x) is the probability that point x is covered,

$$E[\mu(X)] = \int_{-\infty}^{+\infty} p(x)dx$$

Extending to obtain higher moments of $\mu(X)$, we suppose $p(x_1, \ldots, x_m)$ is the probability that $x_1,\ldots,x_m$ all belong to the covered set X. Then:

$$E[\mu(X)^m] = \int_{-\infty}^{+\infty} \ldots\ldots \int_{-\infty}^{+\infty} p(x_1,\ldots\ldots,x_m)dx_1\ldots\ldots dx_m$$

A limitation of the deterministic models is that they assume that servers are available when requested, which is not always true in practical situations. In non-congested systems, with little demand, the assumption is reasonable, but in congested systems, in which frequent calls for service may for example keep ambulances busy 20–30% of the time, the assumption is totally unjustifiable. Congestion in emergency services, which may cause the unavailability of servers within the critical distance when a call is placed, lead to the development of probabilistic covering models.

The first wave of published location models were deterministic and thus did not account for the probability that a particular ambulance might be busy at a given time. As a result, they either underestimated the number of ambulances needed, or overestimated the actual coverage provided. Probabilistic models, on the other hand, acknowledge the

possibility that a given ambulance may not be available when called. This uncertainty has been modeled using queuing or simulation, or embedded into a mathematical programming formulation. [25]

5.3.5 Mathematical programming models

Important models that treat the availability of servers as a random variable are the Maximum Expected Covering Location Problem (MEXCLP) and the Maximum Availability Location Problem (MALP). In both models simplifying assumptions lead to the definition of mathematical programming models: the assumption that servers operate independently is a common feature to both models. In Maximum Expected Covering Location model it is assumed that each server has the same busy probability. As far as it concerns MALP, there are two variations: MALPI, where each server has the same busy probability, and MALPII, where busy fractions are different in the various sections of a region under consideration.

The MEXCLP and MALPI models are both probabilistic extensions of the Maximum Covering Location Problem (MCLP), their deterministic equivalent. The location of emergency services has as a common objective, the provision of coverage to demand areas. The notion of coverage implies the definition of a service distance (time), which is the critical distance (time) beyond which a demand area is considered not covered. A demand area is therefore considered covered if it is within a predefined critical distance (S) from at least one of the existing servers (facilities).

MEXCLP's objective is to maximize the expected coverage of all demand areas under consideration. As we mentioned above, it is assumed that servers operate independently and that all servers have the same busy probability (workload) ρ. In this model it is allowed more than one server to be situated in any given location.

Let J be the set of demand areas. The probability that a demand area $j \in J$ is covered by at least one server, given that k servers cover this area within the critical distance S, is given by:

P [at least one server available within S] = (1- P[no server available within S]) = $1- \rho^k$

Let $H_{j,k}$ be a random variable equal to the demand of area j covered by an available server, given that k servers cover this area. If φ_j is the number of calls per day originating in demand area j, $H_{j,k} = \varphi_j$ with probability $(1 - \rho^k)$, $H_{j,k} = 0$ with probability ρ^k. The expected value of $H_{j,k}$ is given by $E(H_{j,k}) = \varphi_j (1 - \rho^k) \ \forall$ j, k. The increase in expected coverage in demand area j when the number of servers that cover it is increased from (k – 1) to k is given by $\Delta E(H_{j,k}) = E(H_{j,k}) - E(H_{j,k-1}) = \varphi_j \rho^{k-1} (1 - \rho)$, k = 1,2,.....,p.

Let I (|I|= n) be the set of locations where servers may be stationed, $\alpha_{ij} = 1$ if demand area j is covered by a server located at $i \in I$ within critical distance S ($\alpha_{ij} = 0$ otherwise), p is the number of servers to be located. Define variable $x_{jk} \in \{0,1\}$ such that $x_{jk} = 1$ if demand area j has at least k servers within S, $x_{jk} = 0$ otherwise. Finally let $y_i = 0,1,2,.....,p$ represent the number of servers located at $i \in I$. Using the definitions above MEXCLP may be formulated as an integer-programming problem:

$$\text{Maximize} \quad Z = \sum_{j \in J} \sum_{k=1}^{p} \varphi_j (1-\rho) \rho^{k-1} x_{jk} \qquad (1)$$

Subject to:

$$\sum_{i \in I} \alpha_{ij} y_i \geq \sum_{k=1}^{p} x_{jk} \qquad j \in J \qquad (2)$$

$$\sum_{i \in I} y_i \leq p \qquad (3)$$

$$y_i = 0,1,2,.....p \qquad i \in I \qquad (4)$$

$x_{jk} \in \{0,1\} \quad \forall\ j, k$ (5)

In the formulation above the objective function maximizes the expected coverage, considering that up to p servers may cover any given demand area within S. Restrictions (2) count the number of servers that cover demand area j within S, for all $j \in J$. Constraint (3) sets at p the upper limit on the number of servers to be located and restrictions (4) and (5) define the nature of the decision variables. Notice that restrictions (4) allow up to p servers to be situated in any given location $i \in I$. Finally, as the objective function is concave in k for each $j \in J$, it is not necessary to include in this formulation precedence constraints of the type $x_{jk} \leq x_{j(k-1)}$. [12]

As we mentioned above, there are two important probabilistic formulations known as the maximum availability location problems I and II (MALP I & II). Both distribute a fixed number of response units in order to maximize the population covered within a response–time standard and with a predetermined reliability. In MALP I, a system-wide busy probability is computed for all units (similar to MEXCLP), while in MALP II, the region is divided into neighborhoods. Local busy fractions for units in each neighborhood are computed, assuming that the immediate area of interest is isolated from the rest of the region.

Maximum availability location problems I, uses a formula to estimate ρ, the common busy fraction for all servers. Let $\bar{t}$ be the mean service time, measured in hours, for a call originating in any demand area $j \in J$. It is possible to calculate ρ as $\rho = \bar{t} \sum_{j \in J} \varphi_j /24p$, i.e. the busy fraction of each server is calculated dividing the mean number of daily hours of service needed by the system by the number of daily hours available, assuming that p servers will be located. [12]

The restriction that at least one server must be available within S for any given demand area $j \in J$ with probability greater than or equal to α may be written in the following way: P [at least one server available within S] $\geq$ = (1 − P [no server available within S]) $\geq \alpha$ $= 1 - \rho^{\sum_{i \in I} \alpha_{ij} y_i} \geq \alpha$, where $\sum_{i \in I} \alpha_{ij} y_i$ is the number of servers available within S of demand area $j \in J$. Or, taking logarithms, $\sum_{i \in I} \alpha_{ij} y_i \geq d$, where $d = [\log(1 - \alpha)/\log \rho]$.

From the equivalent linear expression obtained from the probabilistic constraint, it is possible to notice that each demand area $j \in J$ requires at least d servers available within critical distance S for it to be covered with reliability α. In order to be able to maximize the number of calls serviced with reliability α, it is necessary therefore to maximize the number of calls with at least d servers available within S.

If we use variable x_{jk} utilized in the definition of MEXCLP, the expression $\sum_{k=1}^{n} x_{jk}$ represents the number of times demand area $j \in J$ is covered within S. In order to maximize the number of calls covered with reliability α, $\sum_{j \in J} \varphi_j x_{jd}$ must be maximized. The mathematical formulation of MALPI can be finally written in the following way:

$$\text{Maximize} \quad Z = \sum_{j \in J} \varphi_j x_{jd} \qquad (6)$$

Subject to:

$$\sum_{k=1}^{d} x_{jk} \leq \sum_{i \in I} \alpha_{ij} y_i \quad j \in J \qquad (7)$$

$$x_{jk} \leq x_{j(k-1)} \quad j \in J,\ k = 2,\ldots\ldots d \qquad (8)$$

$$\sum_{i \in I} y_i = p \qquad (9)$$

$$x_{jk} \in \{0,1\},\ j \in J,\ k = 1,\ldots\ldots d \qquad (10)$$

$y_i \in \{0,1\}, i \in I$ (11)

$1 - \rho^{\sum_{i \in I} \alpha_{ij} y_i} \geq \alpha$ which can be linearized as

$\sum_{i \in I} \alpha_{ij} y_i \geq d$, where $d = [\log(1 - \alpha)/\log \rho]$ (12)

where

$\sum_{i \in I} \alpha_{ij} y_i$: the number of servers available within S of demand area $j \in J$

Restrictions (7) guarantee that a demand area $j \in J$ is covered with reliability α if at least d servers are available within S from j. Constraints (8) express that, for a given demand area to be covered by k servers, it must be covered by at least $(k-1)$ servers, for $2 \leq k \leq d$. These constraints, which omitted in MEXCLP, are necessary in MALPI. Restriction (9) establishes that p servers must be located, while constraints (10) and (11) define the binary nature of the decision variables. Notice that in the formulation of MALPI, only one server may be situated in any location $i \in I$. The model remains valid, however, if more than one server is allowed in any location $i \in I$, as in MEXCLP. In this case constraints (11) would have to be replaced by constraints (4). Finally constraint (12) expresses the minimum number of servers required to serve each demand point j with a reliability level of α. [12]

The formulation above is similar to that of MEXCLP, except of course for the objective function. Notice also the difference between constraints (2) and (7): while in MALPI d servers within S are needed to provide reliability α for demand area j, up to p servers may provide coverage to any demand area $j \in J$ and contribute to the expected coverage expressed in the objective function of MEXCLP.

To make these models mathematically tractable, both employed two simplifying assumptions: (1) servers operate independently, and (2) all servers have the same busy

probability. These assumptions do not reflect the ''real world'' accurately when servers cooperate through centralized dispatching, and they have varying busy probabilities.

5.3.6 Queueing models

The models and methodologies described previously incorporate a range of stochastic problem parameters. In this section, we will see how the probability distributions associated with these parameters have been combined with results from queueing theory to examine additional aspects of facility location. The most well known queuing models for emergency service location problems are the hypercube and approximated hypercube by Larson in 1974 and 1975 [23, 24] which consider the congestions of the system by calculating the steady-state busy fractions of servers on a network. The hypercube model can be used to evaluate a wide variety of output performance such as vehicle utilization, average travel time, inter-district service performance, etc.

Queueing theory, the theory of congestion, is the branch of operations research which explores the relationships between demand on a service system and the delays suffered by the users of that system. Since almost all urban service systems can be viewed as queueing systems, queueing theory plays a central role in the analysis of urban services.

5.3.6.1 Hypercube queueing model

Larson's hypercube model was the first to embedded queueing theory in facility location problems. The model analyzes problems of vehicle location and allocation and response district design in emergency response services that operate in the server-to-customer mode (such as police, fire, emergency medical vehicles). Hypercube model is a descriptive model that must be embedded in an optimization framework in order to search for good solutions.

The solution of the model is the state probabilities and associated system performance measures (workloads, travel times etc). The model considers spatial and temporal complexities of the region and is based on Markovian analysis and queueing theory.

The area's geography is modeled by partitioning a network into a set of geographical nodes, each representing an independent point source of requests for service. A server's primary response area (district) consists of those nodes to which the server would be dispatched if all other servers are available (districting). Each server can be busy or free, generating 2^N possible states for the system (where N is the number of servers). These are the vertices of a hypercube, named B_j ($j = 0, 1, \ldots, 2^{N-1}$) of dimension N. Figure 5-5 represents 5-dimensional hypercube. Each vertex, or state, is denoted by an ordered set of N one digit binary numbers taking the value of 1 if the server is busy and of 0 if not, $B_j \equiv \{b_N, b_{N-1}, \ldots, b_1\}$.

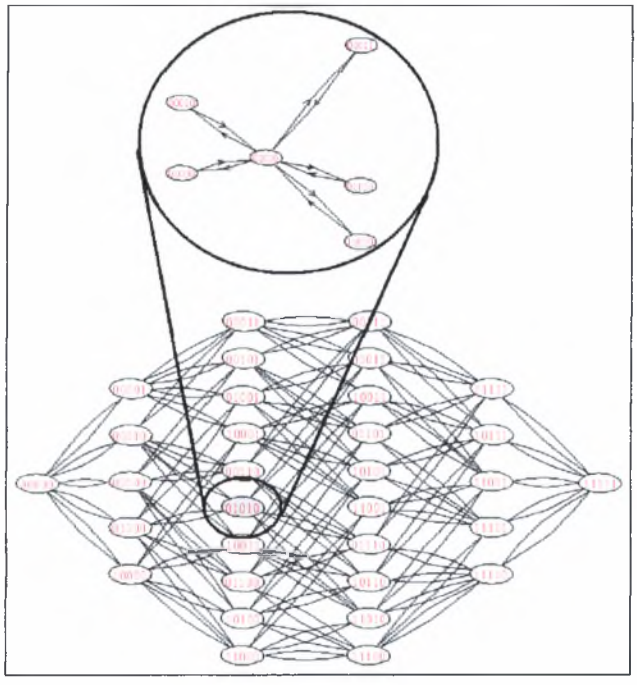


Figure 5-5: Graphical representation of Hypercube with 5 servers.

The HQM assumes that only one step transitions occur and all multistep transitions are not allowable. In other words, transitions are allowed between states with Hamming distance equal to 1, where the Hamming distance d_{ij} between two vertices B_i and B_j is the number of digits by which the two vertices differ (e.g. states 00110 and 00111 have distance equal to 1 and states 01010 and 10100 have distance equal to 4). Larson introduced "upward" and "downward" hamming distance, d_{ij}^{+} and d_{ij}^{-}, as the number of binary digits switching from 0 to 1 and 1 to 0. The geographical depiction of the "location" of a response system is general enough to model fixed or mobile locations of units. This is accomplished by specifying a stochastic (sum of rows equal to 1) location matrix $L = (l_{nj})$, where l_{nj} is the fraction of available time that response unit n spends in node j.

The model assumes Poisson input (requests for service) and exponential service rates, ignoring any past system history. Reasonable deviations from this assumption have been found not to alter the predictive accuracy of the model. The model is thus a finite-state continuous time Markov process whose steady-state probabilities are determined from the equations of detailed balance that express a conservation of flow between consequent states. The location of servers and dispatch preferences is predetermined, i.e., for each geographical tract, an ordered list of units that specifies the dispatcher's preference for units to assign to this tract is known. The model uses the closest available server policy, implying that each time an incoming call is received, the first available server of the list is assigned.

5.3.6.2 Some Uses of the Model

The hypercube queueing model represents an important planning tool that can be used in a variety of applications by planners and administrators of service agencies operating in the

server-to-customer mode. Thus, the following applications are likely to be important in many cities and towns:

Police-sector design. Suppose that a city's police department has not redesigned its sectors for many years. Then, owing to changing population patterns and other factors in the evolution of the city, the distribution of crimes and other incidents that give rise to calls for police service are likely to have changed significantly from the time of last beat design. This could result in an intolerable situation in which some patrol officers are working considerably longer hours responding to calls than are others. Compounding the problem, crime preventive patrol is probably least prevalent in the high-workload areas, since the high call-for-service workload in these areas sharply reduces the time available for patrol.

In this case the model can be used to assist the police planner in redesigning sectors to correct the current imbalances. The model provides outputs on travel times, workloads of each police vehicle, preventive patrol frequencies, and other factors that allow simultaneous consideration of response-time reduction, workload balancing, preventive patrol strength, and so on. The model reveals the trade-offs one must confront in attempting to reach acceptable performance in each of these categories.

In using the model the police planner must specify the sector configuration that he or she desires. Then the model computes the numerical values for each of the performance measures (e.g., travel times, etc.). Undoubtedly, each police planner will have his or her own set of issues-some quantitatively oriented and some not-that will be important in the sector-design process. In most cases, however, regardless of the planner's particular set of issues and their relative priorities, the model described here should be useful in his or her thinking primarily because it computes rapidly and effectively many operationally oriented performance measures that come into play in the sector-design process.

Response-area design for ambulances or emergency repair vehicles. Suppose that an agency administrator disperses ambulances or emergency repair vehicles throughout the city, prepositioning them in a way that best anticipates likely calls for emergency service. Then, the system planner needs assistance in determining good locations for the units and reasonable areas of primary responsibility for each. The model can be used for this purpose, in much the same way as a police planner would use it to design police beats. Here, however, the positions of the vehicles (while not responding to emergencies) are most likely fixed at preselected sites, whereas the police cars are likely to patrol throughout their sectors. The time for an ambulance to service a medical emergency usually includes travel time to and from a hospital (to transport the patient), a time not experienced in the police example (except when transporting arrestees to a police stationhouse). So in the ambulance case it is much more likely that travel times (time to the scene, time from the scene to the hospital, time from the hospital back to the prepositioning site) will play a dominant role in the overall time required per incident. This effect should be less prevalent in the emergency repair case. (In the police case, on-scene service time is usually significantly greater than travel time.)

Thus, the ambulance or emergency repair system planner can use the model to explore the consequences of alternative prepositioning sites for his vehicles and alternative districts of primary responsibility for each. Since travel times play such an important role in ambulance services, it is likely that the emergency medical planner will have to adjust the service time of each ambulance separately to reflect the different geographical travel time factors affecting each one. The final site selection and district design could include factors of workload balance, travel-time reduction, neighborhood integrity, and so on. Again, analogous to the police-sector example, the exact trade-off among the various factors must be determined by the user of the model, not by the model itself. [13, 25]

Chapter 6 Applications of Emergency Response Vehicles Location Models – Three Cases

6.1 An application of HQM

In this chapter, we will study some applications of emergency response vehicles location models for a profound comprehension of them. Firstly, an application of hypercube queueing model and secondly an application of P-Median model.

We consider a three-server city shown in Figure 6-1. This city is partitioned into 10 geographical tracts, each acting as an independent Poisson generator of service requests. For convenience, we assume that the mean service time of each unit n is the same known constant μ^{-1}.

The rates λ_i, of arrival of requests from each tract i are shown in Figure 6-1 (expressed in arrivals per mean service time unit). Each unit has a primary response area, consisting of a set of tracts to which it would always be given first dispatch preference. For instance, the primary response area for unit 2 consists of tracts 3, 5, and 6. The unit given second dispatch preference for a tract is selected on the basis of geographical proximity. The complete dispatch preference policy, by tract, is shown in Table 6-1. Note, for instance, that unit 1 is the primary backup unit for all service requests in both primary response areas 2 and 3. Thus, not only does unit 1 face a heavy workload from its own primary response area (50 percent of the city's workload), but it is also the first backup unit for the rest of the city as well.

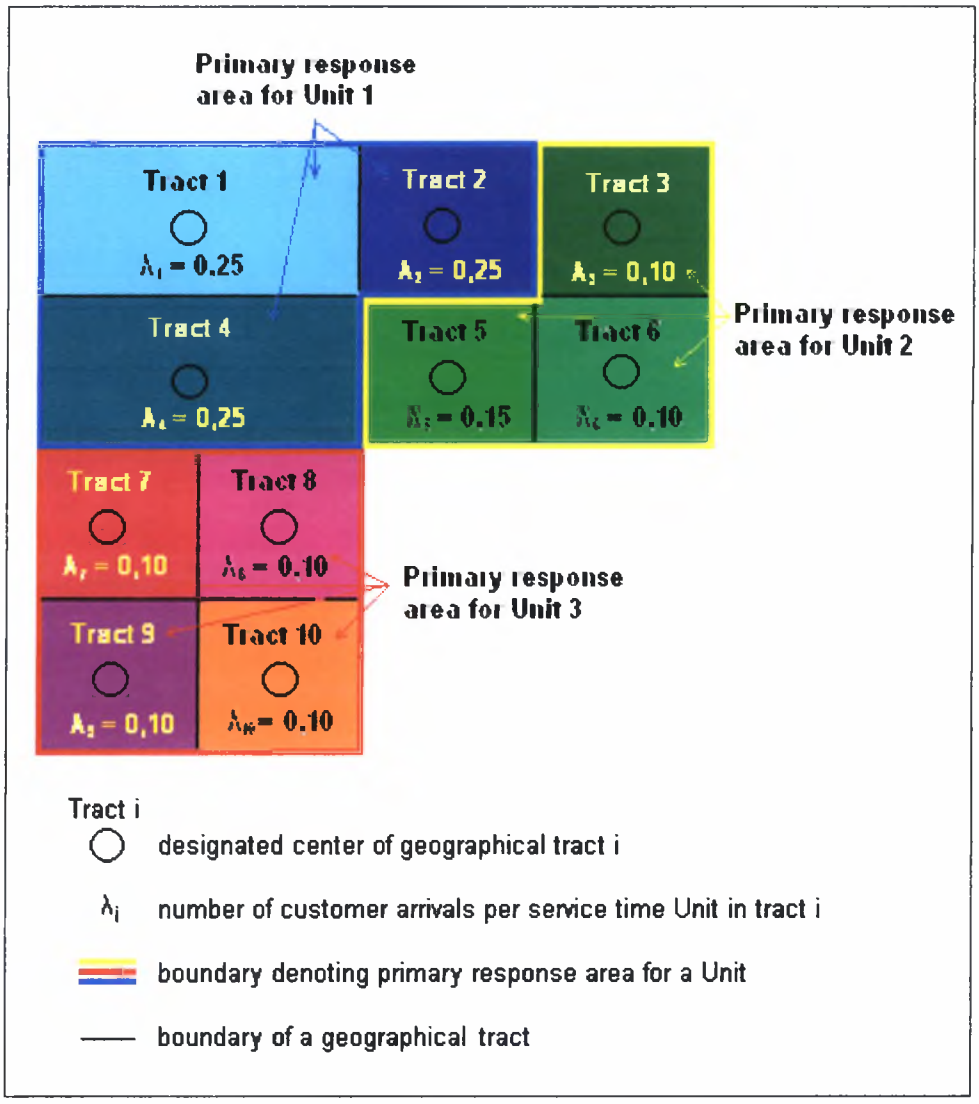


Figure 6-1: Map of three-server city.

In our example we have set all the mean service times equal to the same constant. With this restriction, the queueing system in the aggregate is simply the M / M / N system. As we mentioned above, if we say that b_n = state of server n (b_n = 0 for "free," or b_n = 1 for "busy"), the system state is given by B = $\{b_N, b_{N-1}, ..., b_1\}$The collection of all possible B's is denoted by C_N, corresponding to the vertices of an N-dimensional hypercube. The weight of

B, denoted w(B), is equal to the number of busy servers in the state B, thus, $w(B) = \sum_{n=1}^{N} b_n$.

For instance, the weight of {0, 1, 1} is 2 and the weight of {0, 0, 0} is 0.

Tract Number	First Preference Unit	Second Preference Unit	Third Preference Unit
1	1	2	3
2	1	2	3
3	2	1	3
4	1	3	2
5	2	1	3
6	2	1	3
7	3	1	2
8	3	1	2
9	3	1	2
10	3	1	2

Table 6-1: Dispatch preferences for three-server city.

For the case of equal mean service times, we can use the concept of weight of a state to relate the hypercube state space to the simpler M / M / N state space. This equivalence is obtained by collecting together all states having equal weight, i.e. w_0, their summed probability of occurrence is equal to the comparable probability of state S_{w0}, occurring in the M / M / N model. For instance, the states 001, 010, and 100 all have weight 1, and their summed probability must equal the probability of state S_1, in the M / M / N model. One way of demonstrating this equivalence is shown in Figure 6-2, in which all hypercube states having equal weight are grouped together vertically. Figure 6-2 also, shows explicitly the infinite tail that augments the hypercube state space to allow the possibility of a queue.

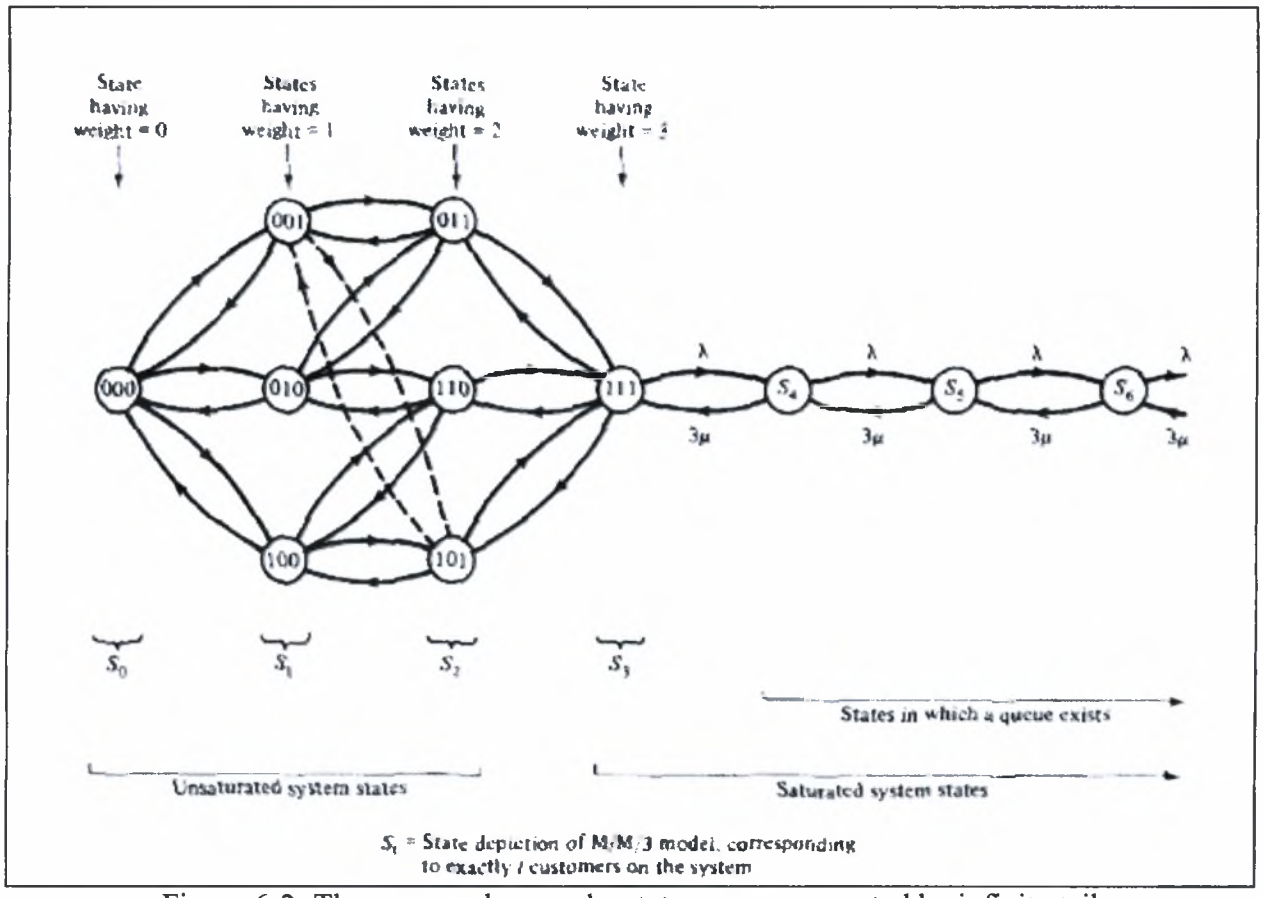


Figure 6-2: Three-server hypercube state space augmented by infinite tail.

With the hypercube model transitions occur between states just as they do with other queuing models. The hypercube model assumes that only one unit is assigned to each service request. Thus, assuming that it is virtually impossible for two requests to arrive simultaneously, a transition from state (0, 0, 0) to state (0, 1, 0) would be allowable, whereas a two-step transition from state (0, 0, 0) to state (0, 1, 1) would not (since it implies that two free units become busy at the same instant). Likewise, if we assume that each of the busy units is working independently at its particular location to complete service on its current service request, it is virtually impossible for two busy units to become free simultaneously. Thus, a transition from state (1, 0, 1) to state (0, 0, 1) is allowable, whereas a two-step transition from state (1, 0, 1) to state (0, 0, 0) is not.

In our example where N = 3 unit problem, we suppose that the following events occur:

1. A request for service arrives from tract 6.
2. A request for service arrives from tract 1.
3. Unit 2 completes service on its request.
4. A request for service arrives from tract 4.
5. A request for service arrives from tract 3.

Recalling that tract 6 is within unit 2's primary response area, the request for service from tract 6 would result in unit 2 becoming busy, yielding a transition from state (0, 0, 0) to state (0, 1, 0). Likewise, the next request for service would cause unit 1 to become busy, resulting in a transition from state (0, 1, 0) to state (0, 1, 1). Next, when unit 2 becomes free, the system makes a transition to state (0, 0, 1). Next, when a request for service arrives from tract 4, in unit 1's primary response area, we note that unit 1 is already busy servicing a request. Since the primary backup unit, unit 3, is available, it is dispatched to the scene, resulting in an interresponse area assignment. The system now undergoes a transition from state (0, 0, 1) to state (1, 0, 1). Finally, when a request arrives from tract 3, unit 2 is dispatched, causing the system to enter a saturation state (1, 1, 1). (Any additional service requests arriving while all servers are busy would be delayed in queue.) Summarizing the example above, the sequence of states occupied by the system is as follows:

1. State where no unit is busy → (0,0,0)
2. A request for service arrives from tract 6 → (0,1,0)
3. A request for service arrives from tract 1 → (0,1,1)
4. Unit 2 completes service on its request → (0,0,1)
5. A request for service arrives from tract 4 → (1,0,1)
6. A request for service arrives from tract 3 → (1,1,1)

Now the value of the cube as an aid in visualizing the behavior of the system is seen in Figure 6-3, which depicts the states occupied and the state-to-state transitions as a sequence of connected trips along edges of the cube.

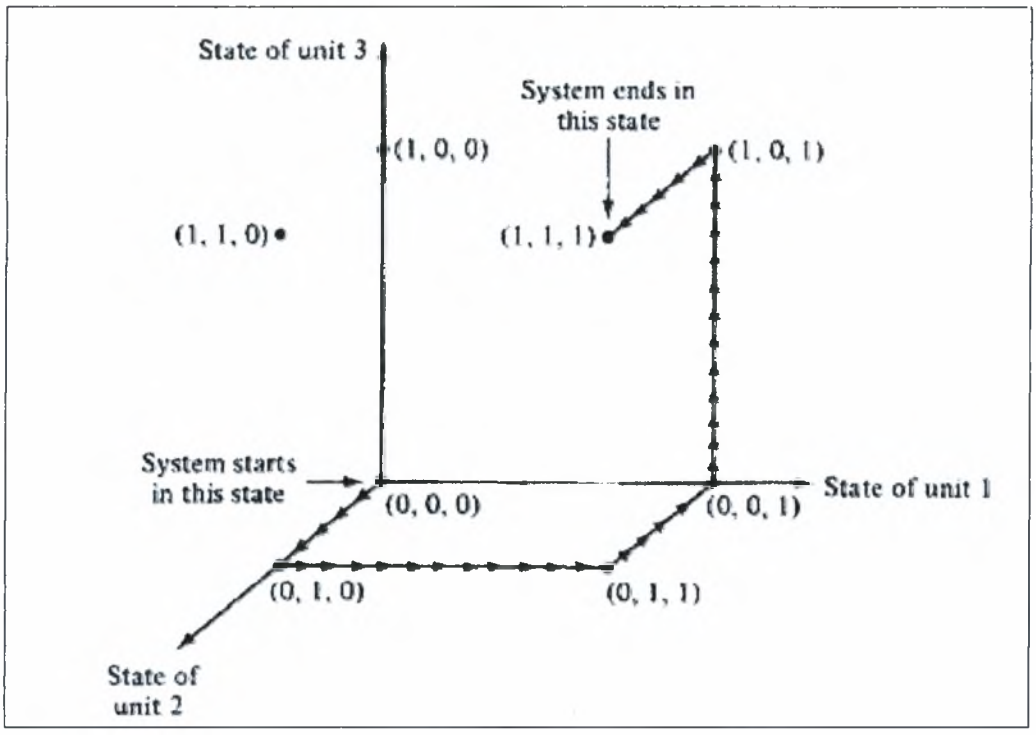


Figure 6-3: Illustrative sequence of transitions.

Now, we link the arrival rates and the service rates to states and transitions on the hypercube. Recall that 0,75 request for service arrives per unit time from unit 1's primary response area, 0,35 from unit 2's primary response area, and 0,40 from unit 3's. Also recall that time is measured in mean service time units. Then, from state (0, 0, 0), there is a rate of transition to state (0, 0, 1) equal to 0,75 request per unit time. Likewise, there is a rate of transition from state (0, 0, 0) to state (0, 1, 0) equal to 0,35 per hour and to state (1, 0, 0) equal to 0,40 per unit time. These state-to-state transition rates can be drawn onto the cube as shown in Figure 6-4. In a similar manner, the transition rate from any state to any adjacent state having one less unit busy is 1 per unit time, these transition rates are depicted in Figure 6-5.

We have now filled in the transition rates from state (0, 0, 0) and all the rates corresponding to completions of service. For convenience, the latter type of transitions are called "downward" transitions, indicating that the total number of busy units has dropped

down by one. Transitions that result in a unit being dispatched are called "upward" transitions, because the total number of busy units has gone up by one.

Since all service requests that arrive are serviced immediately (i.e., they incur no queue delay) if at least one response unit is available, all states that are unit distance from the saturation state (states 011, 101, and 110 in our example) must have an upward transition rate equaling the total system-wide request rate λ (= 1.5 in our case). Thus, all upward transition rates into state 111 must equal 1.5 in our example.

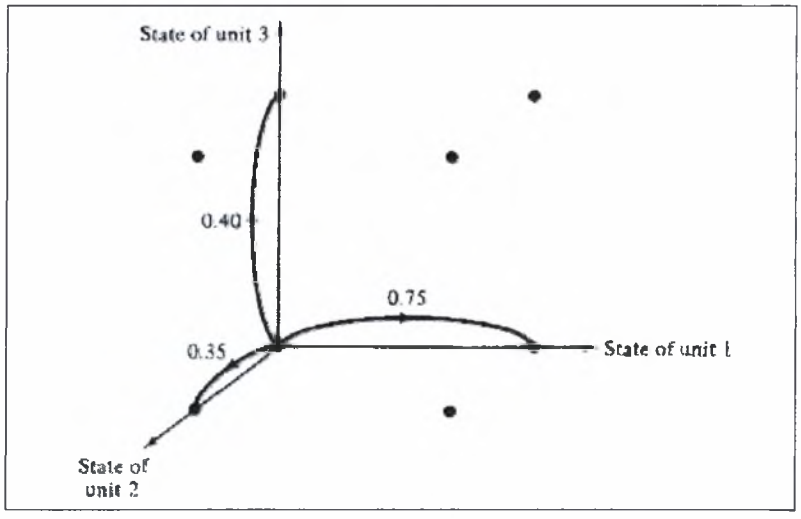


Figure 6-4: Transition rates of state (0,0,0).

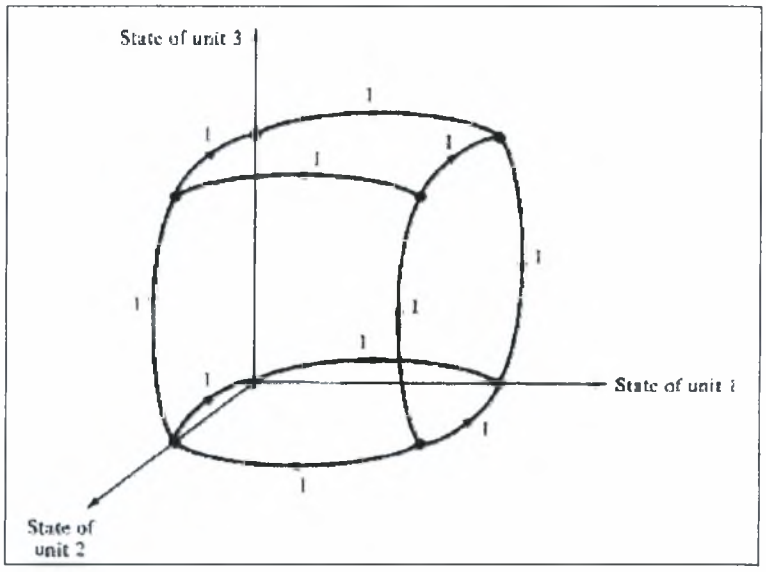


Figure 6-5: Transition rates corresponding to completions of service.

To compute the values of the remaining upward transition rates, consider for instance the transition rate from state 001 to state 101. This rate will consist of the sum of two rates: the rate of service requests from unit 3's primary response area plus the overflow rate from that part of unit 1's primary response area assigned to unit 3 as the first backup unit. The first rate is simply 0,40 request per unit time. The second is the arrival rate from tract 4 (0,25 request per unit time), since the first backup unit for tract 4 (in unit 1's primary response area) is unit 3. (Unit 2 is the first backup unit for the other tracts in unit 1's primary response area.) Thus, the net upward transition rate from state 001 to state 101 is (0,40 + 0,25) = 0,65 request per unit time. In a like manner, the remaining upward transition rates can be found. The entire state-transition diagram (excluding the infinite tail) is shown in Figure 6-6.

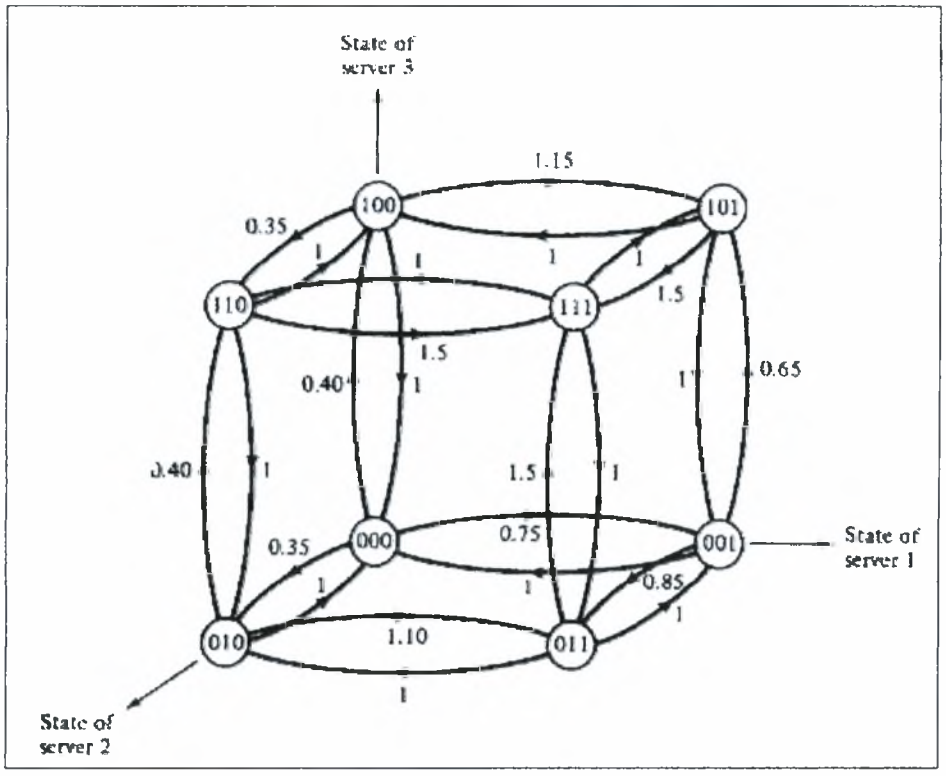


Figure 6-6: Hypercube state-transition diagram for unsaturated system states: three server city (states involving a queue not included).

We define

P_{ijk} = steady-state probability that the system is in state {i, j, k}, i, j, k = 0, 1

$P\{S_i\}$ = steady-state probability that the equivalent M / M / 3 system is in state S_i

P_Q = steady-state probability that a queue of positive length exists

From the M / M / 3 model and the relationships:

$$P_o = \left[\sum_{n=0}^{m-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^m}{m!} \frac{1}{1-(\lambda/m\mu)} \right]^{-1}$$

(assuming that $\lambda/m\mu < 1$)

and

$$P_n = \frac{(\lambda/\mu)^n}{n!} P_o \quad \text{for n = 0,1,....,m-1}$$

$$P_n = \frac{(\lambda/\mu)^n}{m^{n-m} \cdot m!} P_o \quad \text{for n = m,m+1,m+2.....}$$

setting $\lambda = 1{,}5$, $\mu = 1$, we have:

$$P\{S_o\} = \left[1 + 1{,}5 + 1{,}125 + 1{,}125\right]^{-1} = 0{,}2105$$

$$P\{S_1\} = 1{,}5 * P\{S_o\} = 0{,}3157$$

$$P\{S_2\} = 1{,}125 * P\{S_o\} = 0{,}2368$$

$$P\{S_3\} = 0{,}5625 * P\{S_o\} = 0{,}1184$$

$$P_Q = \sum_{i=4}^{\infty} P\{S_i\} = 1 - \sum_{i=0}^{3} P\{S_i\} = 0{,}1186$$

Using these results, we can now write the hypercube equilibrium equations for nonsaturated system states:

1. Empty state

$P_{ooo} = P\{S_o\} = 0{,}2105$

2. Full state (no queue)

$P_{111} = P\{S_3\} = 0{,}1184$

3. First hyperplane from the origin

$P_{001} + P_{010} + P_{100} = P\{S_1\} = 0{,}3157$

4. Second hyperplane from the origin

$P_{110} + P_{101} + P_{011} = P\{S_2\} = 0{,}2368$

5. Balance of flow about state 001

$$P_{001}\{0,65+0,85+1\} = P_{000}\underbrace{\begin{pmatrix}0,75\\ \uparrow \\ \text{flow rate from 000 to 001}\end{pmatrix}}_{\text{total flow out of state 001}} + 1\cdot\underbrace{\begin{pmatrix}\underbrace{P_{101}+P_{011}}_{\text{total downflow into state 001}}\end{pmatrix}}_{\text{total flow into state 001}}$$

6. Balance of flow about state 100

$$P_{100}(1,15+0,35+1) = P_{000}(0,40) + 1\cdot(P_{110}+P_{101})$$

7. Balance of flow about state 011

$$P_{011}(1+1,5+1) = P_{010}(1,10) + P_{111}(1) + P_{001}(0,85)$$

8. Balance of flow about state 101

$$P_{101}(1+1,5+1) = 1\cdot P_{111} + P_{001}(0,65) + P_{100}(1,15)$$

9. Balance of flow about state 010

$$P_{010}(0,40+1+1,10) = 1\cdot P_{110} + P_{000}(0,35) + 1\cdot P_{011}$$

10. Balance of flow about state 110

$$P_{110}(1+1,5+1) = (0,35)\,P_{100} + P_{010}(0,40) + 1\cdot P_{111}$$

We solve this set of equations. After calculations, we arrive at the following values for the state probabilities:

$P_{000} = 0,2105$

$P_{001} = 0,1367$

$P_{010} = 0,0886$

$P_{100} = 0,0905$

$P_{110} = 0,0530$

$P_{101} = 0,0889$

$P_{011} = 0,0949$

$P_{111} = 0,1184$

Now that we know the steady-state probabilities of the hypercube model, we can obtain values of useful system performance measures such as the workloads.

We obtain the workloads of the individual servers. The workload ρ_n of server n, which is the fraction of time that server n is busy, is equal to the sum of all steady-state probabilities having the state of server n equal to 1 (rather than 0) plus the fraction of time that a queue exists (during which time all servers are working). Thus, for our three-server example,

$\rho_1 = P_{001} + P_{101} + P_{011} + P_{111} + P_Q = 0,5574$

$\rho_2 = P_{010} + P_{110} + P_{011} + P_{111} + P_Q = 0,4734$

$\rho_3 = P_{100} + P_{110} + P_{101} + P_{111} + P_Q = 0,4693$

As we can see, these results check with the requirement that the average workload $\rho = \lambda/3\mu = 0,5$. Note that the workload sharing among response units caused the workloads of the units to be more evenly distributed than the workloads of the primary response areas. If each unit served only the customers of its own response area, the workloads would have been $\rho_1 = 0,75$, $\rho_2 = 0,35$, $\rho_3 = 0,40$. In fact, it is possible for a particular primary response area to generate more work than one unit could handle, and workload sharing would facilitate the overflow.

6.1.1 Sensitivity Analysis

Until now, we considered the rates λ_i, of arrival of requests from each tract i as constant values: $\lambda_1 = 0,25$, $\lambda_2 = 0,25$, $\lambda_3 = 0,10$, $\lambda_4 = 0,25$, $\lambda_5 = 0,15$, $\lambda_6 = 0,10$, $\lambda_7 = 0,10$, $\lambda_8 = 0,10$, $\lambda_9 = 0,10$ and $\lambda_{10} = 0,10$. Let's assume that we want to know any time the state probabilities of the system and the workloads of the individual servers changing λ_1, λ_2.....λ_{10} values keeping constant the total rate λ of arrival of requests from all tracts, $\lambda = 1,5$ and the events occurred previously.

In this case, we transform Figure 6-6 in a more general state-transition diagram shown in Figure 6-7.

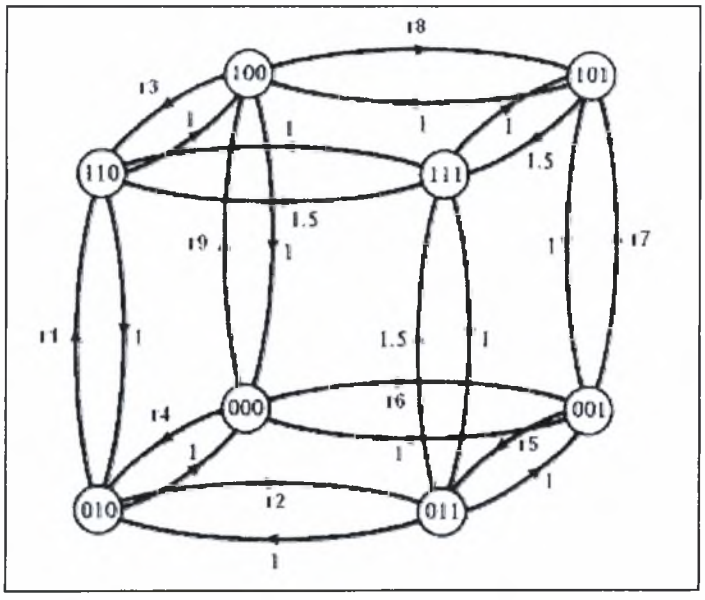


Figure 6-7: A transformed more general state-transition diagram.

Where we define:

$r_1 = \lambda_7+\lambda_8+\lambda_9+\lambda_{10}$

$r_2 = r_4+\lambda_1+\lambda_2+\lambda_4$

$r_3 = \lambda_3+\lambda_5+\lambda_6$

$r_4 = \lambda_3+\lambda_5+\lambda_6$

$r_5 = r_6 + \lambda_6$

$r_6 = \lambda_1+\lambda_2+\lambda_4$

$r_7 = \lambda_7+\lambda_8+\lambda_9+\lambda_{10}+\lambda_4$

$r_8 = r_9+\lambda_1+\lambda_2+\lambda_4$

$r_9 = \lambda_7+\lambda_8+\lambda_9+\lambda_{10}$

Now, if we change the rates λ_i, of arrival of requests from each tract i, rates r_1, r_2,.....r_{10} change, too. From the M / M / 3 model and the relationships:

$$P_o = \left[\sum_{n=0}^{m-1}\frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^m}{m!}\frac{1}{1-(\lambda/m\mu)}\right]^{-1}$$ (assuming that $\lambda/m\mu < 1$)

and

$$P_n = \frac{(\lambda/\mu)^n}{n!}P_o \quad \text{for n = 0,1,....,m-1}$$

$$P_n = \frac{(\lambda/\mu)^n}{m^{n-m}\cdot m!}P_o \quad \text{for n = m,m+1,m+2.....}$$

setting $\lambda = \mathbf{1,5}$, $\mu = 1$, we have:

$P\{S_o\} = [1+1,5+1,125+1,125]^{-1} = 0,2105$

$P\{S_1\} = 1,5 * P\{S_o\} = 0,3157$

$P\{S_2\} = 1,125 * P\{S_o\} = 0,2368$

$P\{S_3\} = 0,5625 * P\{S_o\} = 0,1184$

$$P_Q = \sum_{i=4}^{\infty} P\{S_i\} = 1 - \sum_{i=0}^{3} P\{S_i\} = 0,1186$$

As we notice, $P\{S_o\}$, $P\{S_1\}$, $P\{S_2\}$, $P\{S_3\}$, $P\{S_4\}$ are the same as previously because these probabilities depend only on total rate $\lambda = 1,5$ but the hypercube equilibrium equations for

nonsaturated system states give us new values for state probabilities P_{001}, P_{010}, P_{100}, P_{110}, P_{101}, P_{011} and new workloads of the individual servers. In order to estimate these values we use an Excel spreadsheet as it is shown in Figure 6-8.

It is clear that, for a new set of rates λ_i: $\lambda_1 = 0{,}010$, $\lambda_2 = 0{,}010$, $\lambda_3 = 0{,}005$, $\lambda_4 = 0{,}010$, $\lambda_5 = 0{,}015$, $\lambda_6 = 0{,}010$, $\lambda_7 = 0{,}040$, $\lambda_8 = 0{,}050$, $\lambda_9 = 0{,}050$ and $\lambda_{10} = 1{,}3$ but constant total rate $\lambda = 1{,}5$, we arrive at the following new values for the state probabilities:

$P_{000} = 0{,}2105$ $\quad$ $P_{001} = 0{,}0771$

$P_{010} = 0{,}0371$ $\quad$ $P_{100} = 0{,}2017$

$P_{110} = 0{,}0508$ $\quad$ $P_{101} = 0{,}1505$

$P_{011} = 0{,}0353$ $\quad$ $P_{111} = 0{,}1184$

while the workloads of servers 1,2,3 are $\rho_1 = 0{,}4999$, $\rho_2 = 0{,}3602$, $\rho_3 = 0{,}6399$ respectively.

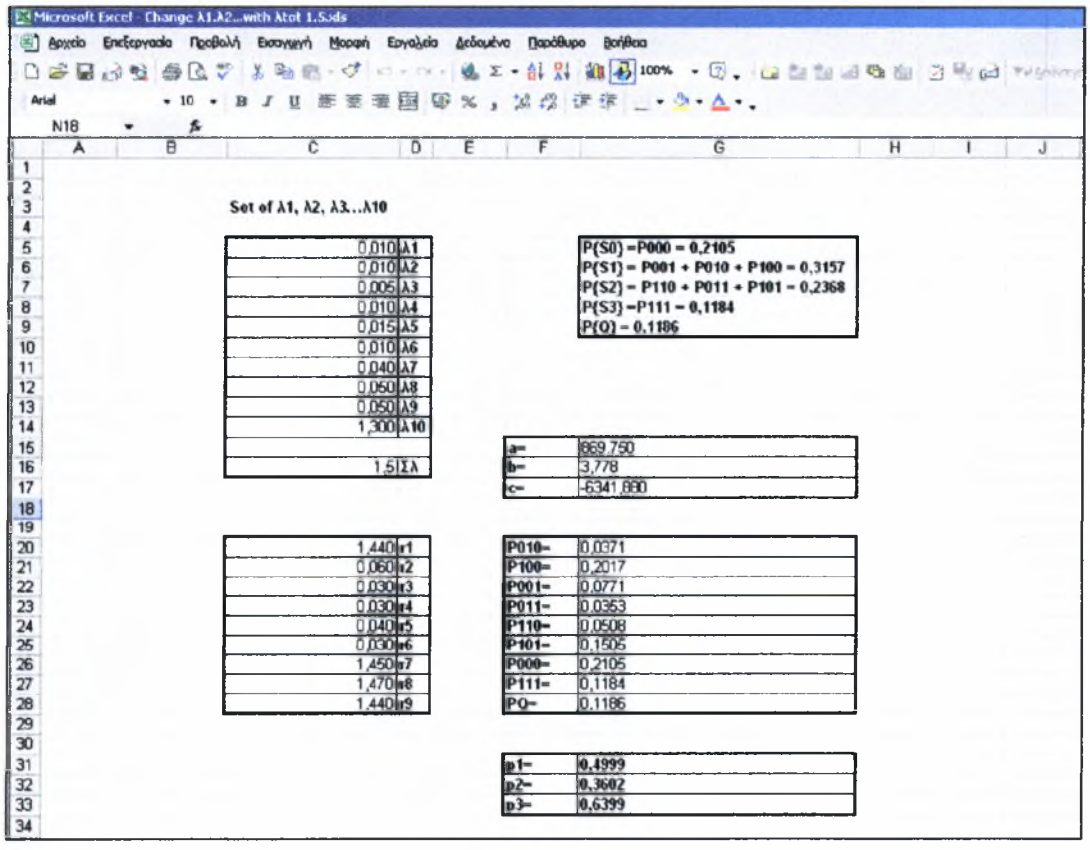


Figure 6-8: Estimation of new state probabilities and servers' workloads for different sets of $\lambda_1, \lambda_2 \ldots \lambda_{10}$.

We notice that, the workloads of the three servers are not evenly distributed for this set of rates. Comparing to the previous set of rates, a high increase of the rate λ_{10} results to a heavy workload of server 3.

So, this Excel spreadsheet, can inform us, any time for different values of λ_1, λ_2.....λ_{10} and constant value of the total rate λ of arrival of requests from all tracts, $\lambda = 1,5$, and the events occurred previously, about the state probabilities of the system and the workloads of the three servers.

Figure 6-9 depicts the fluctuation of the three servers workloads for fifty different random sets of λ_1, λ_2, λ_3.....λ_{10} and constant value of total rate λ with $\Sigma\lambda = 1,5$. It is obvious that the workload of server 1 is over 0,5 for the most of different random sets of λ_1, λ_2, λ_3.....λ_{10} and its average value is 0,51. As far as it concerns the workload of server 2, it remains under 0,5 for quite all sets of λ_1, λ_2, λ_3.....λ_{10} and its average value is 0,47.

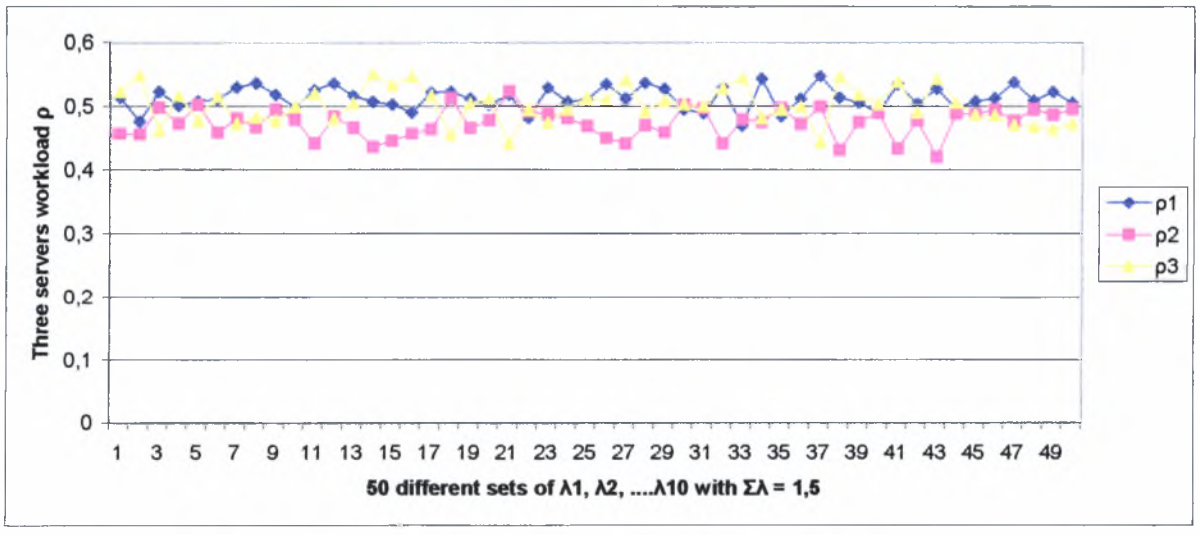


Figure 6-9: Three servers workloads for different random sets of λ_1, λ_2, λ_3.....λ_{10} and constant value of the total rate λ of arrival of requests from all tracts, $\Sigma\lambda = 1,5$.

The average value of workload of server 3 is 0,5 with 50% of values being over 0,5 and 50% of them under 0,5. So, if we increase the number of different random sets of λ_1, λ_2, λ_3.....λ_{10} we notice that the average workload of server 1 is heavier than the workload of server 2 and 3.

A general case of the previous example is the state probabilities and the workloads estimation of the individual servers changing now the total rate λ of arrival of requests from all tracts and λ_1, λ_2.....λ_{10} values, too. We transform Figure 6-6 in its general state-transition diagram shown in Figure 6-10.

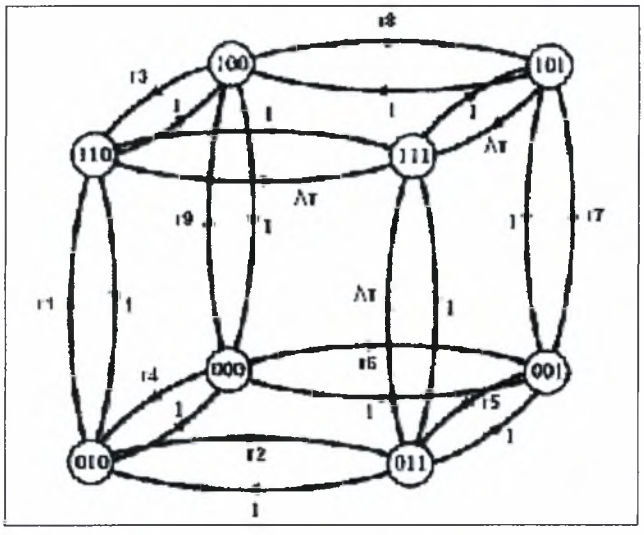


Figure 6-10: A transformed general state-transition diagram.

We define again:

$r_1 = \lambda_7+\lambda_8+\lambda_9+\lambda_{10}$

$r_2 = r_4+\lambda_1+\lambda_2+\lambda_4$

$r_3 = \lambda_3+\lambda_5+\lambda_6$

$r_4 = \lambda_3+\lambda_5+\lambda_6$

$r_5 = r_6 + \lambda_6$

$r_6 = \lambda_1+\lambda_2+\lambda_4$

$r_7 = \lambda_7+\lambda_8+\lambda_9+\lambda_{10}+\lambda_4$

$r_8 = r_9+\lambda_1+\lambda_2+\lambda_4$

$r_9 = \lambda_7+\lambda_8+\lambda_9+\lambda_{10}$

Now let's assume that the total rate $\lambda = 2,5$, the following relationships arise:

$$P_o = \left[\sum_{n=0}^{m-1}\frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^m}{m!}\frac{1}{1-(\lambda/m\mu)}\right]^{-1}$$ (assuming that $\lambda/m\mu < 1$)

and

$$P_n = \frac{(\lambda/\mu)^n}{n!}P_o \quad \text{for n = 0,1,....,m-1}$$

$$P_n = \frac{(\lambda/\mu)^n}{m^{n-m}\cdot m!}P_o \quad \text{for n = m,m+1,m+2.....}$$

setting $\lambda = \mathbf{2,5}$, $\mu = 1$, we have:

$P\{S_o\} = 0,1111$

$P\{S_1\} = 0,1124$

$P\{S_2\} = 0,1404$

$P\{S_3\} = 0,1170$

$$P_Q = \sum_{i=4}^{\infty} P\{S_i\} = 1 - \sum_{i=0}^{3} P\{S_i\} = 0,5852$$

As we notice, $P\{S_o\}$, $P\{S_1\}$, $P\{S_2\}$, $P\{S_3\}$, $P\{S_4\}$ are not the same as previously because of the change of total rate $\lambda = 2,5$. Solving the hypercube equilibrium equations for nonsaturated system states we conclude to the following new values for the state probabilities:

$P_{000} = 0,1111$

$P_{001} = 0,0376$

$P_{010} = 0,0404$

$P_{100} = 0,0343$

$P_{110} = 0,0419$

$P_{101} = 0,0454$

$P_{011} = 0{,}0491$

$P_{111} = 0{,}1170$

$P_Q = 0{,}5852$

Now, we change the value of the total rate λ of arrival of requests and we study the influence of $\Sigma\lambda$ to workloads of server 1, 2 and 3. The results for different values of $\Sigma\lambda$ are illustrated in Table 6-2 and Figure 6-11.

$\Sigma\lambda$	ρ_1	ρ_2	ρ_3
1,5	0,507117	0,463608	0,516574
1,7	0,568366	0,545732	0,577573
1,9	0,660782	0,608722	0,636887
2,1	0,72242	0,711669	0,658025
2,4	0,795152	0,786025	0,81308
2,5	0,842845	0,818348	0,837445

Table 6-2 : Workloads ρ_1, ρ_2 and ρ_3 for different values of total rate λ of arrival of requests

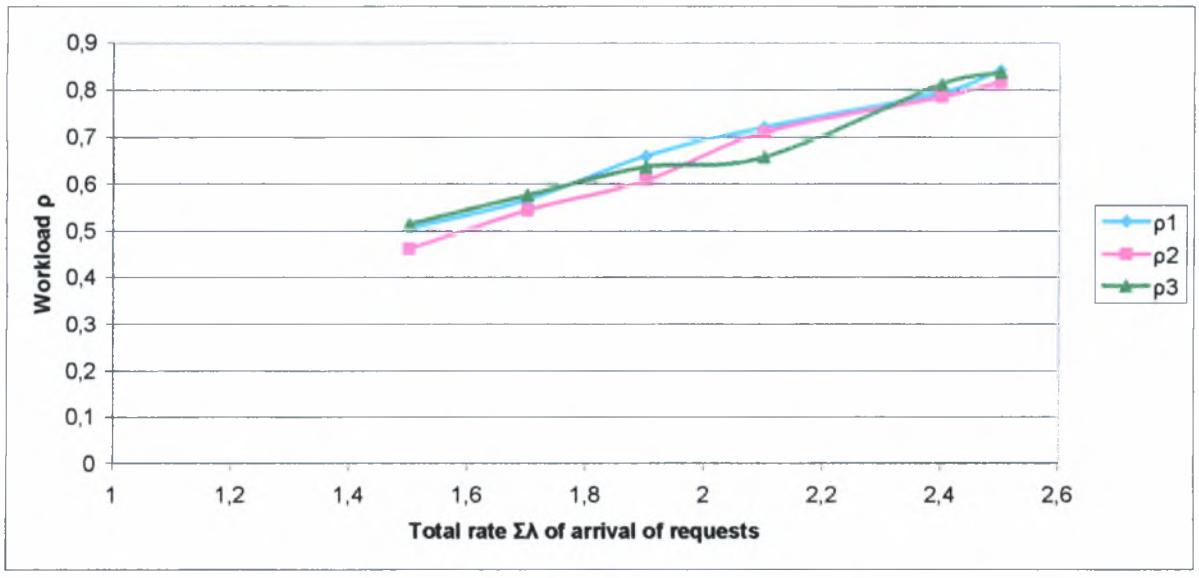


Figure 6-11: Workloads fluctuation for different total rate $\Sigma\lambda$ of arrival of requests

Initally, we see that if Σλ varies from 1,5 to 1,8 server 3 has the heavier workload and follows the workload of server 1 and server 2. If Σλ varies from 1,8 to 2, we have a workloads increase of all servers. Server 1 now has the heavier workload and follows the workload of server 3 and server 2. If Σλ is between 2 and 2,3 then server 1 has still the heavier workload, follows the workload of server 2 and last the workload of server 3 and at the same time all three workloads increase gradually. Finally, if Σλ varies from 2,3 to 2,5, server 3 has again the heavier workload as in the first part of the graph and follows the workload of server 1 and last the workload of server 2. Generally, it is obvious that as Σλ increases the three workloads tend to be equal.

Furthermore, an increase of the total rate λ of arrival of requests from all tracts from 1,5 to 2,5 can lead to an increase in the probability that a queue of positive length exists.

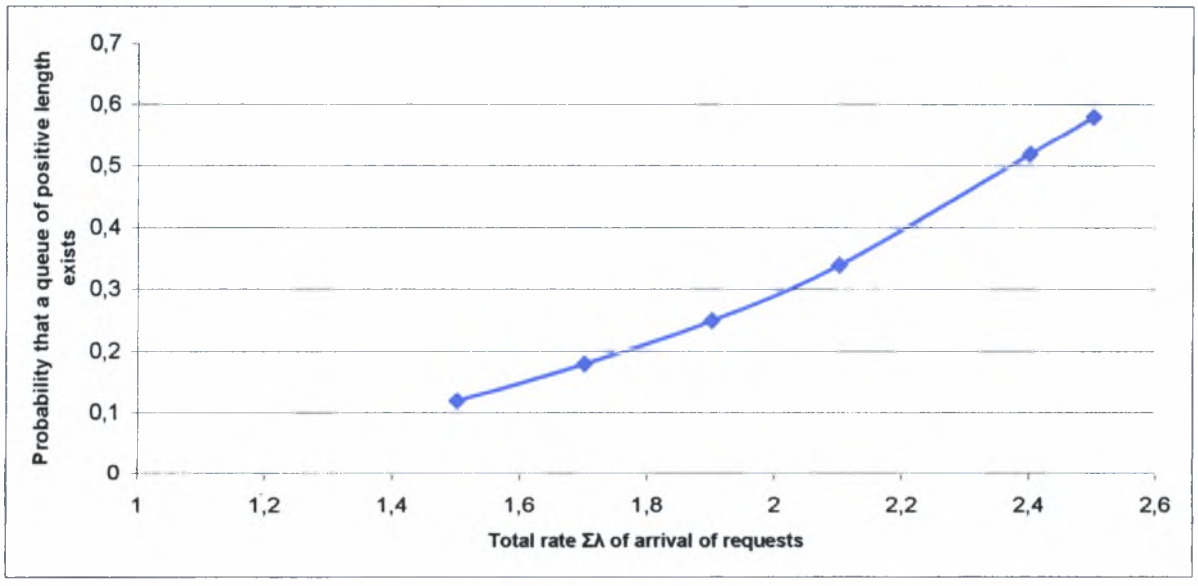


Figure 6-12: Probability queue for different values of total rate λ of arrival of requests

Once again, in order to estimate these values for several λ, we use an Excel spreadsheet as it is shown in Figures 6-13 and 6-14.

So, this Excel spreadsheet, can inform us, any time for different values of total rate λ of arrival of requests from all tracts and λ_1, λ_2.....λ_{10} values, keeping constant the events occurred previously, about the state probabilities of the system, the workloads of the three servers and the probability that a queue of positive length exists.

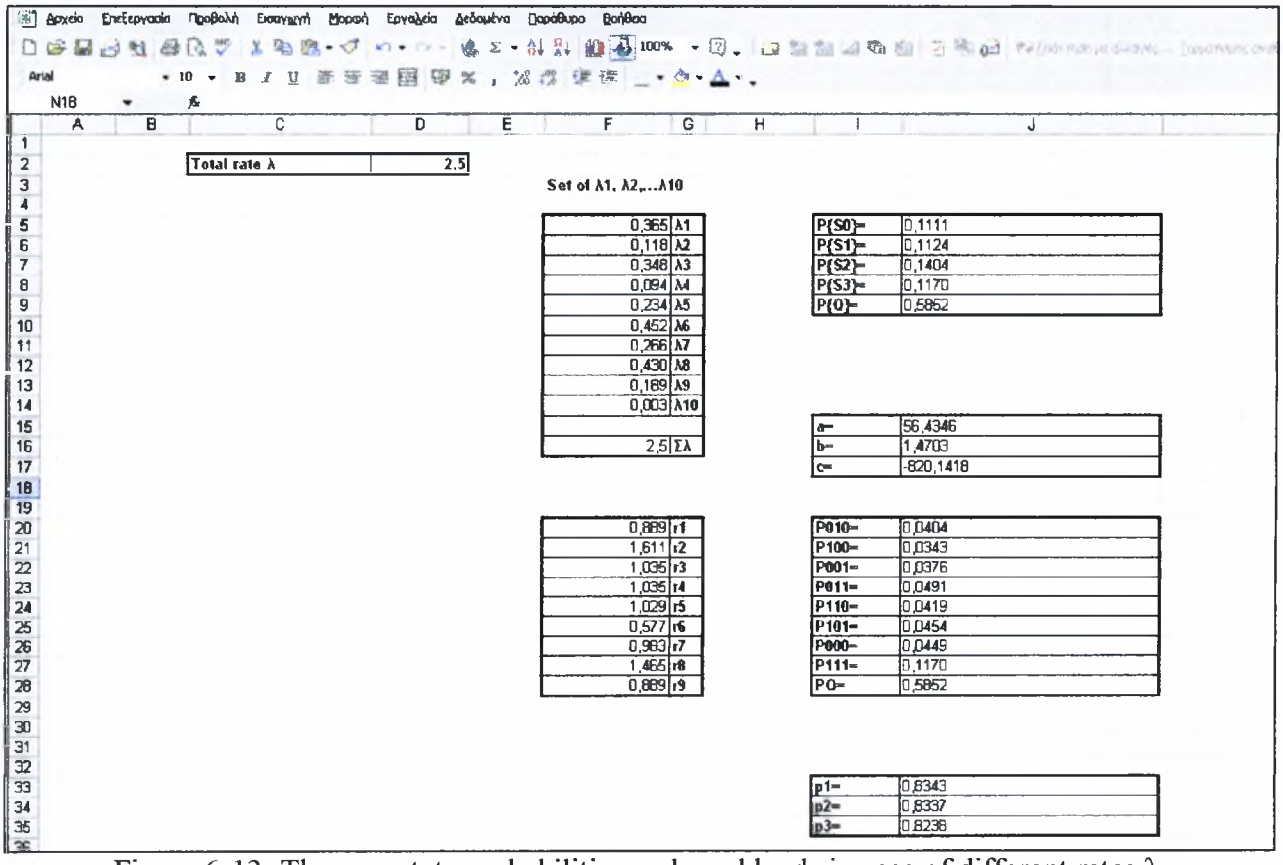
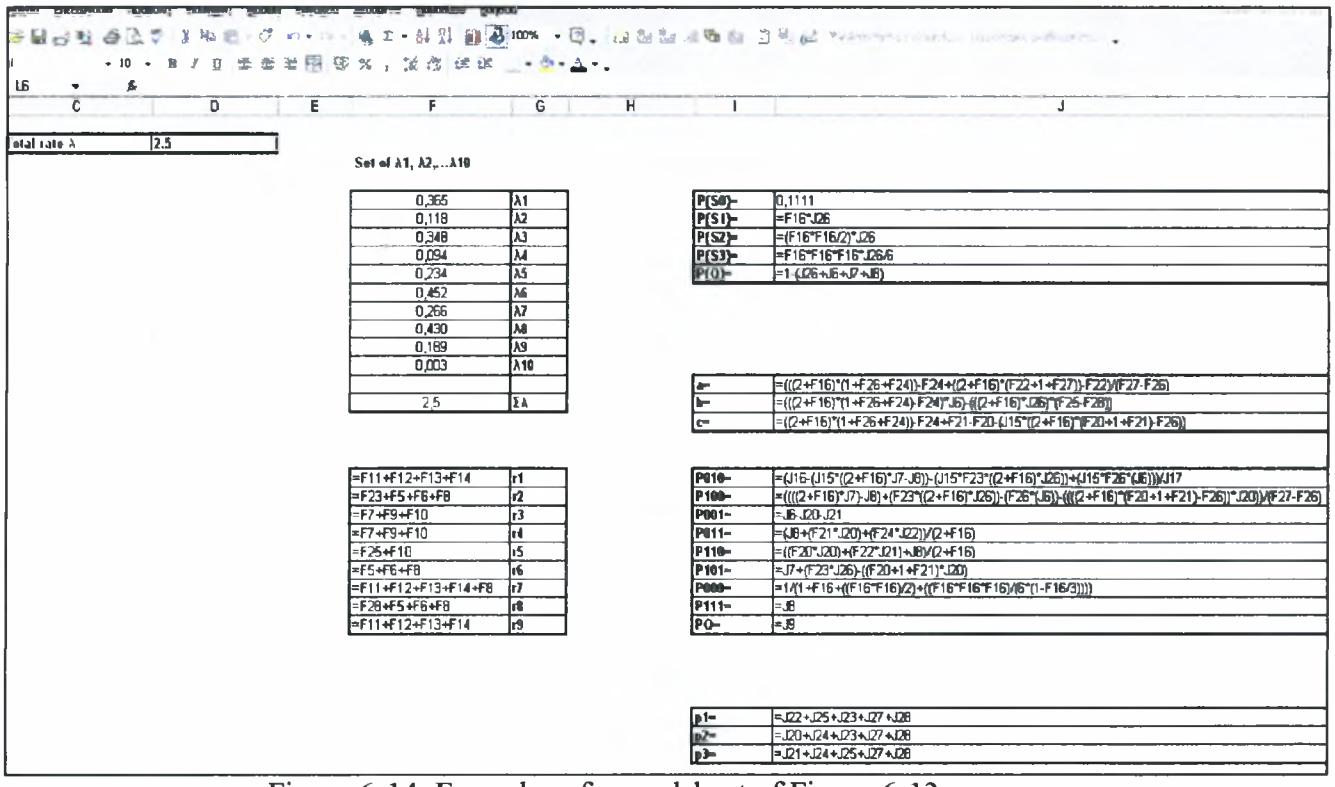


Figure 6-13: The new state probabilities and workloads in case of different rates λ_i.



	C	D	E	F	G	H	I	J
	total rate λ	2,5						
				Set of λ1, λ2,...λ10				
				0,365	λ1			
				0,118	λ2			
				0,348	λ3			
				0,094	λ4			
				0,234	λ5			
				0,452	λ6			
				0,266	λ7			
				0,430	λ8			
				0,189	λ9			
				0,003	λ10			
				2,5	Σλ			
				=F11+F12+F13+F14	r1			
				=F23+F5+F6+F8	r2			
				=F7+F9+F10	r3			
				=F7+F9+F10	r4			
				=F25+F10	r5			
				=F5+F6+F8	r6			
				=F11+F12+F13+F14+F8	r7			
				=F28+F5+F6+F8	r8			
				=F11+F12+F13+F14	r9			
							P(S0)=	0,1111
							P(S1)=	=F16*J26
							P(S2)=	=(F16*F16/2)*J26
							P(S3)=	=F16*F16*F16*J26/6
							P(0)=	=1-(J26+J6+J7+J8)
							a=	=(((2+F16)*(1+F26+F24)-F24+((2+F16)*(F22+1+F27))-F22)/(F27-F26)
							b=	=(((2+F16)*(1+F26+F24)-F24)*J6)-(((2+F16)*J26)*(F25-F26))
							c=	=((2+F16)*(1+F26+F24))-F24+F21-F20-(J15*((2+F16)*(F20+1+F21)-F26))
							P010=	=(J16-(J15*((2+F16)*J7-J6))-(J15*F23*((2+F16)*J26))+(J15*F26*(J6)))/J17
							P100=	=((((2+F16)*J7)-J8)+(F23*((2+F16)*J26))-(F26*(J6))-(((2+F16)*(F20+1+F21)-F26))*J20))/(F27-F26)
							P001=	=J6-J20-J21
							P011=	=(J8+(F21*J20)+(F24*J22))/(2+F16)
							P110=	=((F20*J20)+(F22*J21)+J8)/(2+F16)
							P101=	=J7+(F23*J26)-((F20+1+F21)*J20)
							P000=	=1/(1+F16+(((F16*F16)/2)+(((F16*F16*F16)/6*(1-F16/3))))
							P111=	=J8
							PQ=	=J8
							p1=	=J22+J25+J23+J27+J28
							p2=	=J20+J24+J23+J27+J28
							p3=	=J21+J24+J25+J27+J28

Figure 6-14: Formulas of spreadsheet of Figure 6-13.

Considering data from the example described previously, shown in Figure 6-1, unit 1 has a primary response area which consists of tracts 1, 2, 4. Now, we keep rates λ_2, ...λ_{10} of arrival of requests constant and we examine how the workloads of the three units change for different values of λ_1. The results are illustrated in the following Table 6-3 and the graph in Figure 6-15.

λ_1	ρ_1	ρ_2	ρ_3
0,25	0,5575	0,4734	0,4694
0,5	0,5981	0,4780	0,4693
1	0,6706	0,4958	0,4693
1,2	0,6973	0,5051	0,4693
1,5	0,7356	0,5209	0,4695

Table 6-3: Workloads ρ_1, ρ_2 and ρ_3 for different values of rate λ_1 of arrival of requests of tract 1.

We notice that, an increase of λ_1 results to a heavier workload of unit 1 while workload of unit 2 increases faintly and workload of unit 3 remains at the same level.

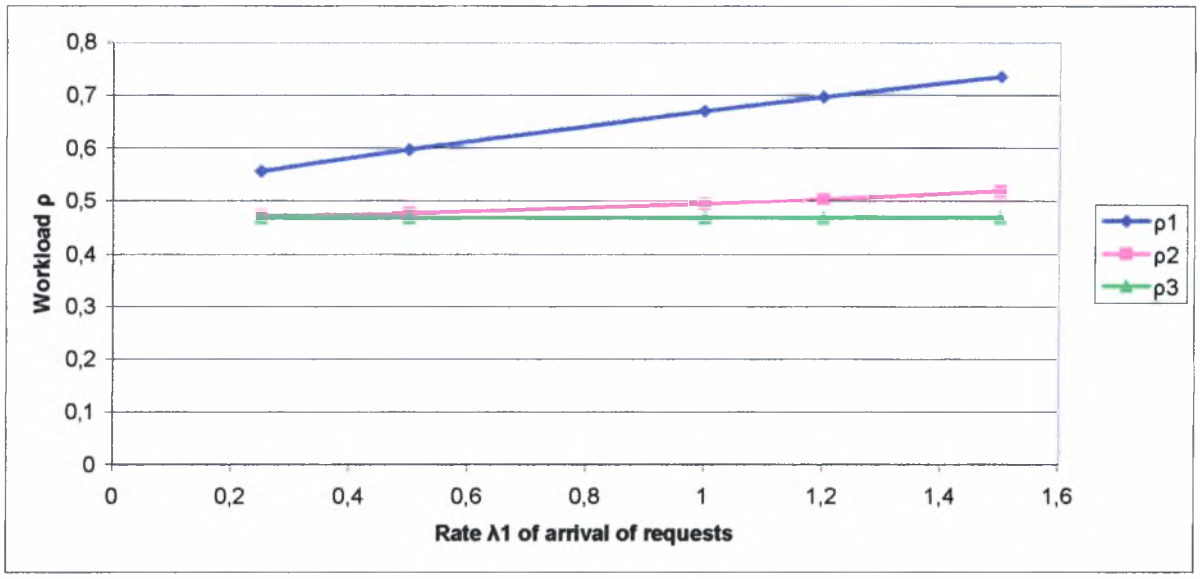


Figure 6-15: Workloads ρ_1, ρ_2, ρ_3 depending on rate λ_1 of arrival of requests.

Now, we keep rates $\lambda_1,\ldots,\lambda_3$ and $\lambda_5,\ldots,\lambda_{10}$ of arrival of requests constant and we examine how the workloads of the three units change for different values of λ_4. The results are illustrated in the following Table 6-4 and the graph in Figure 6-16.

λ_4	ρ_1	ρ_2	ρ_3
0,08	0,5251752	0,477071	0,455187
0,15	0,5386408	0,475457	0,460805
0,25	0,5575	0,4735	0,4694
1,3	0,7414927	0,468942	0,581849
1,7	0,6735411	0,468059	0,536006

Table 6-4: Workloads ρ_1, ρ_2 and ρ_3 for different values of rate λ_4 of arrival of requests of tract 4.

We notice that, an increase of λ_4 results to a heavier workload of unit 1 and unit 3 for $\lambda_4 \leq 1,4$. If $\lambda_4 > 1,4$, workload of unit 1 and 3 decreases, respectively. Furthermore, workload of unit 2 remains quite constant for any increment of λ_4.

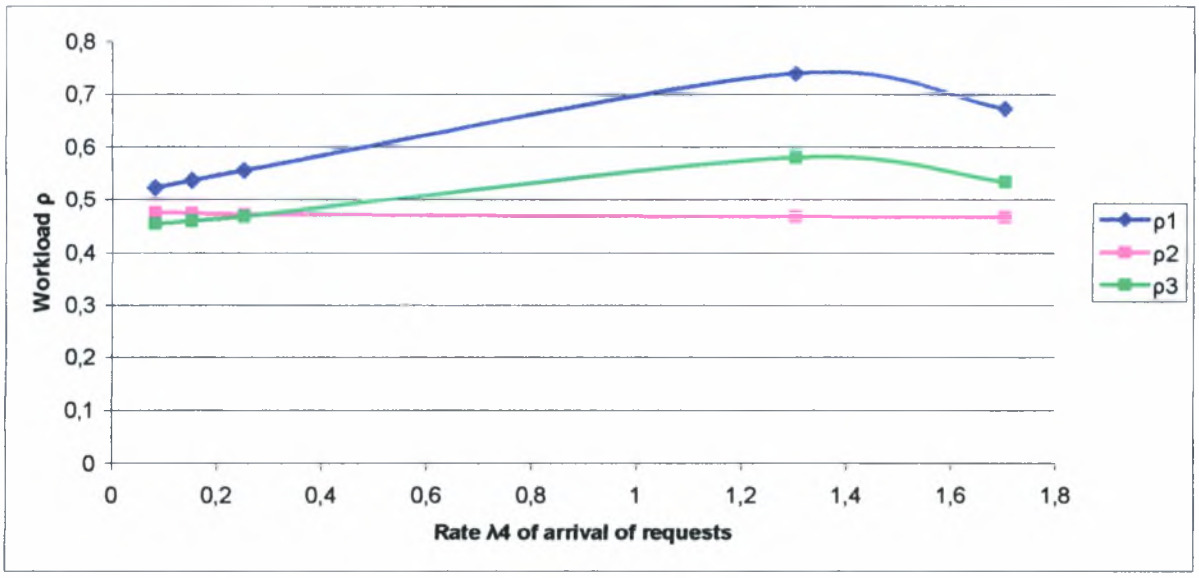


Figure 6-16: Workloads ρ_1, ρ_2, ρ_3 depending on rate λ_4 of arrival of requests

As far as it concerns unit 2 has a primary response area which consists of tracts 3, 5, 6. Now, we keep rates λ_1, ...λ_5 and λ_7...λ_{10} of arrival of requests constant and we examine how the workloads of the three units change for different values of λ_6. The results are illustrated in the following Table 6-5 and the graph in Figure 6-17.

λ_6	ρ_1	ρ_2	ρ_3
0,1	0,5575	0,4735	0,4694
0,75	0,5872	0,5667	0,4637
1,3	0,6205	0,6355	0,4609
1,5	0,6337	0,6591	0,4602
1,6	0,6404	0,6707	0,4599

Table 6-5: Workloads ρ_1, ρ_2 and ρ_3 for different values of rate λ_6 of arrival of requests of tract 6.

We notice that, an increase of λ_6 results to a heavier workload of unit 2 while workload of unit 1 increases faintly and workload of unit 3 remains at the same level. If $\lambda_6 \leq 1$, unit 1 faces the heavier workload. If $\lambda_6 > 1$, workload of unit 2 is heavier than workload of unit 1.

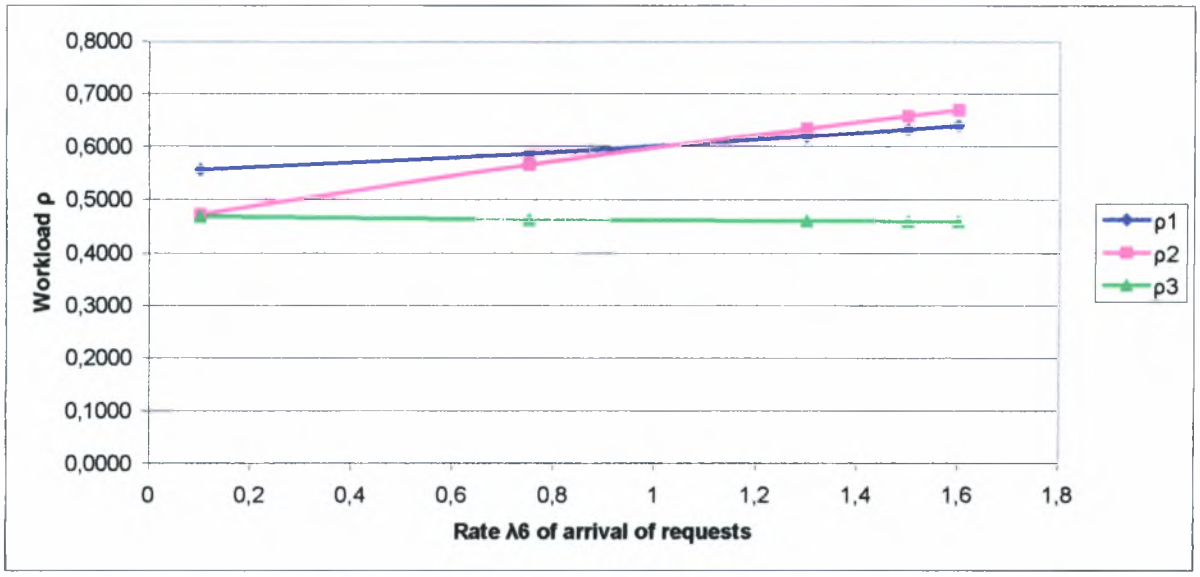


Figure 6-17: Workloads ρ_1, ρ_2, ρ_3 depending on rate λ_6 of arrival of requests.

Now, we keep rates $\lambda_1,...\lambda_4$ and λ_6, $...\lambda_{10}$ of arrival of requests constant and we examine how the workloads of the three units change for different values of λ_5. The results are illustrated in the following Table 6-6 and the graph in Figure 6-18.

λ_5	ρ_1	ρ_2	ρ_3
0,15	0,5575	0,4735	0,4694
0,35	0,56179732	0,497208	0,462038
0,75	0,57454856	0,539183	0,448836
1,1	0,58881744	0,571546	0,43843
1,35	0,60021591	0,592804	0,431434

Table 6-6: Workloads ρ_1, ρ_2 and ρ_3 for different values of rate λ_5 of arrival of requests of tract 5.

We notice that, an increase of λ_5 results to a little heavier workload of unit 1 while workload of unit 2 increases significantly and workload of unit 3 decreases.

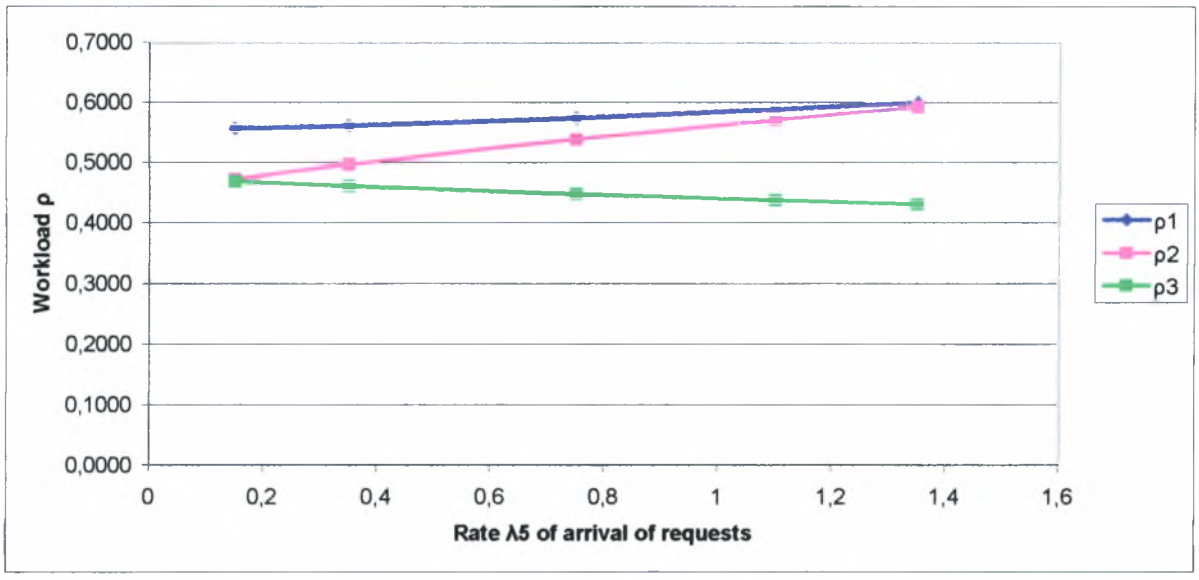


Figure 6-18: Workloads ρ_1, ρ_2, ρ_3 depending on rate λ_5 of arrival of requests

Unit 3 has a primary response area which consists of tracts 7, 8, 9 and 10. Now, we keep rates λ_1, ...λ_7 and λ_9 and λ_{10} of arrival of requests constant and we examine how the workloads of the three units change for different values of λ_8. The results are illustrated in the following Table 6-7 and the graph in Figure 6-19.

λ_8	ρ_1	ρ_2	ρ_3
0,1	0,5575	0,4735	0,4694
0,5	0,5779	0,4457	0,5489
0,95	0,6106	0,4183	0,6249
1,4	0,6496	0,3934	0,6919
1,6	0,6684	0,3830	0,7196

Table 6-7: Workloads ρ_1, ρ_2 and ρ_3 for different values of rate λ_8 of arrival of requests of tract 8.

We notice that, an increase of λ_8 results to a heavier workload of unit 3 while workload of unit 1 increases faintly and workload of unit 2 decreases significantly. If $\lambda_8 \leq 0{,}8$, unit 1 faces the heavier workload. If $\lambda_8 > 0{,}8$, workload of unit 3 is heavier than workload of unit 1.

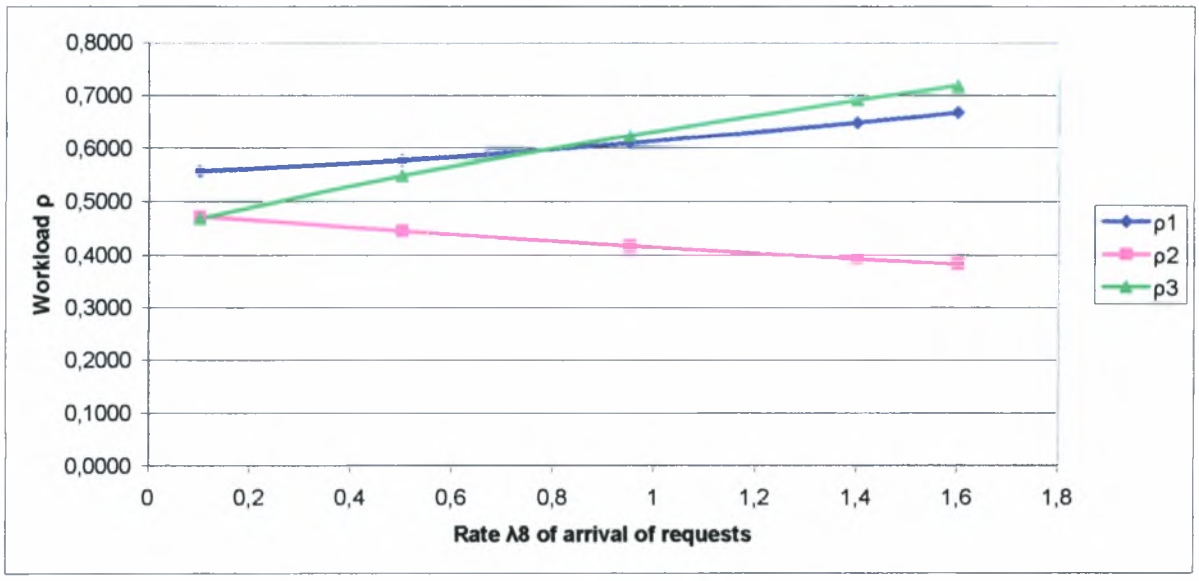


Figure 6-19: Workloads ρ_1, ρ_2, ρ_3 depending on rate λ_8 of arrival of requests

Now, we keep rates λ_1, ...λ_9 of arrival of requests constant and we examine how the workloads of the three units change for different values of λ_{10}. The results are illustrated in the following Table 6-8 and the graph in Figure 6-20.

λ_{10}	ρ_1	ρ_2	ρ_3
0,1	0,5575	0,4735	0,4694
0,65	0,5879	0,4362	0,5755
1,25	0,6360	0,4015	0,6703
1,65	0,6732	0,3804	0,7264
1,8	0,6878	0,3729	0,7464

Table 6-8: Workloads ρ_1, ρ_2 and ρ_3 for different values of rate λ_{10} of arrival of requests of tract 10.

We notice that, an increase of λ_{10} results to a significant heavier workload of unit 1 and 3 while workload of unit 2 decreases.

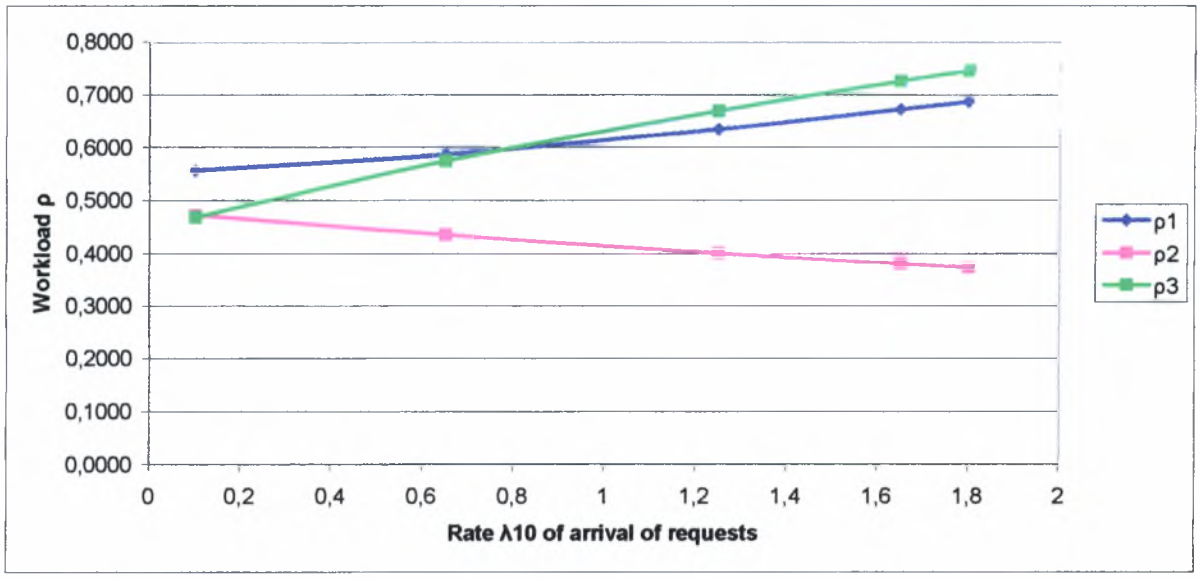


Figure 6-20: Workloads ρ_1, ρ_2, ρ_3 depending on rate λ_{10} of arrival of requests

6.2 An application of P-Median problem

A representative application of this model is the following example. Suppose that authorities of a new city have to decide where to allocate two fire mobile stations (vehicles) in the city. The city has been divided into five tracts as illustrated in Figure 6-21, with no more than one fire station to be located in any given tract. Each station is to respond to all of the fires that occur in the tract in which it is located, as well as in the other tracts that are assigned to the station.

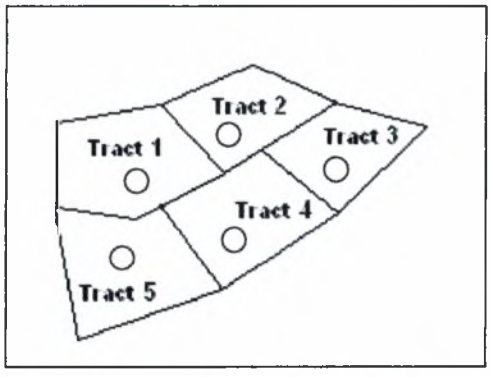


Figure 6-21: The city partitioned into 5 tracts.

The following Table 6-9 gives the average response time to a fire in each tract if their tract is served by a station in a given tract. The bottom row gives the forecasted average number of fires that will occur in each of tracts each day. The objective is to minimize the overall average of the response times to fires.

Station/Fire	Tract 1	Tract 2	Tract 3	Tract 4	Tract 5
Station 1	5 min	12 min	30 min	20 min	15 min
Station 2	20 min	4 min	15 min	10 min	25 min
Station 3	15 min	20 min	6 min	15 min	12 min
Station 4	25 min	15 min	25 min	4 min	10 min
Station 5	10 min	25 min	15 min	12 min	15 min
Frequency	**2**	**1**	**3**	**1**	**3**

Table 6-9: Average response time to a fire in each tract (columns) if their tract is served by a station in a given tract (rows).

We define :

s, t: indices for tract T = {1,2,...5}

d(s, t): response time/ distance between s and t

f(t): frequency of fire in tract t

$$x(s,t) = \begin{cases} 1 & \text{Fire in tract 't' will be covered by station in tract 's'} \\ 0 & \text{otherwise} \end{cases}$$

$$y(s) = \begin{cases} 1 & \text{Build a fire station in tract 's'} \\ 0 & \text{otherwise} \end{cases}$$

and the model can be formulated as follows:

$$\text{Minimize } Z = \sum_{s\in T}\sum_{t\in T} x(s,t)\cdot d(s,t)\cdot f(t)$$

Subject to:

$$\sum_{s\in T} y_s = 2$$

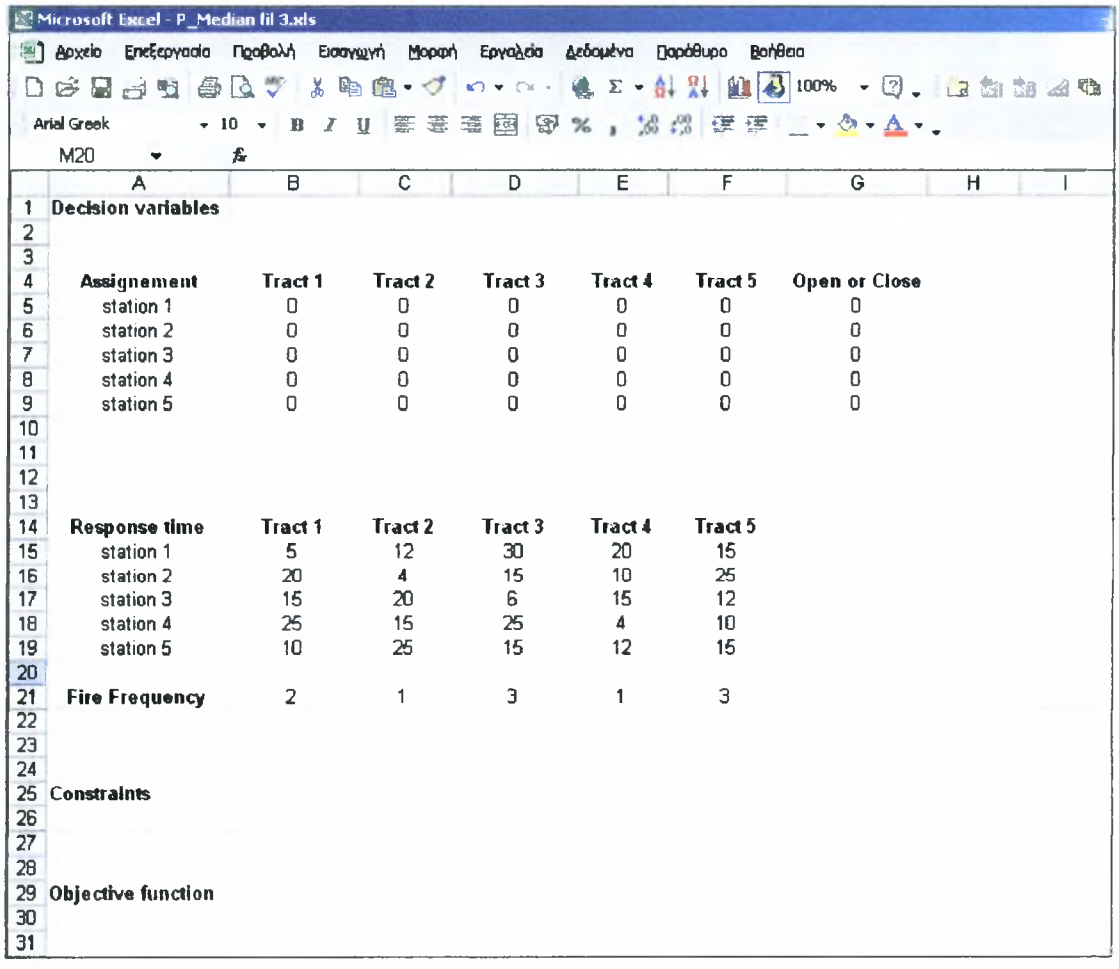
$$\sum_{s\in T} x(s,t) = 1 \qquad \forall t \in T$$

$$x(s,t) \le y_s \qquad \forall s,t \in T$$

$$x(s,t),\ y(t) \in \{0,1\} \quad \forall s,t \in T$$

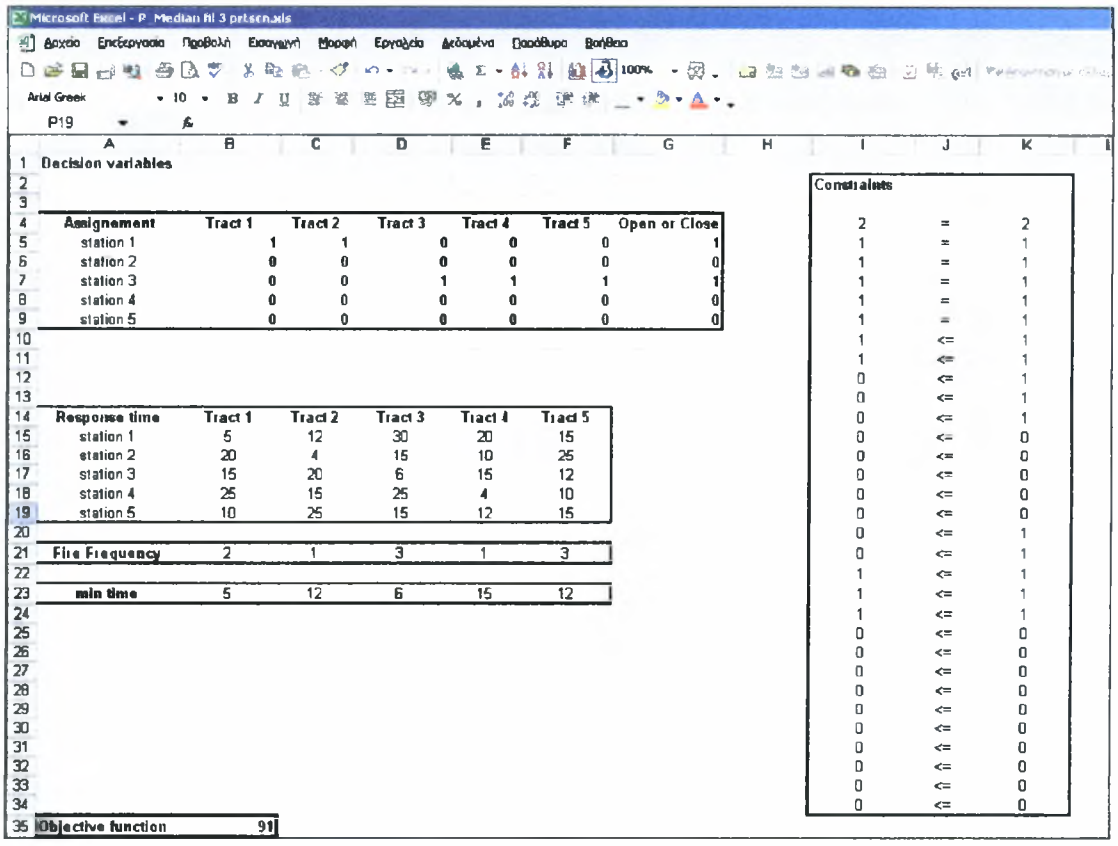
The objective function minimizes the overall average response time to fires. The first constraint states that we are going to locate exactly 2 fire stations. The second constraint states that each tract has to be covered by a fire station so that some one will respond to the fire. The third constraint states that demand tracts can only be assigned to open fire stations and the forth constraint states that no more than one station can be allocated in each tract.

This problem can be solved either using Excel Solver or P-Median Algorithm. If we use the Excel Solver, the solution is depicted in Figures 6-22 and 6-23.



	A	B	C	D	E	F	G	H	I
1	**Decision variables**								
2									
3									
4	**Assignement**	**Tract 1**	**Tract 2**	**Tract 3**	**Tract 4**	**Tract 5**	**Open or Close**		
5	station 1	0	0	0	0	0	0		
6	station 2	0	0	0	0	0	0		
7	station 3	0	0	0	0	0	0		
8	station 4	0	0	0	0	0	0		
9	station 5	0	0	0	0	0	0		
10									
11									
12									
13									
14	**Response time**	**Tract 1**	**Tract 2**	**Tract 3**	**Tract 4**	**Tract 5**			
15	station 1	5	12	30	20	15			
16	station 2	20	4	15	10	25			
17	station 3	15	20	6	15	12			
18	station 4	25	15	25	4	10			
19	station 5	10	25	15	12	15			
20									
21	**Fire Frequency**	2	1	3	1	3			
22									
23									
24									
25	**Constraints**								
26									
27									
28									
29	**Objective function**								
30									
31									

Figure 6-22: Excel solver solution (1).



Decision variables

Assignement	Tract 1	Tract 2	Tract 3	Tract 4	Tract 5	Open or Close
station 1	1	1	0	0	0	1
station 2	0	0	0	0	0	0
station 3	0	0	1	1	1	1
station 4	0	0	0	0	0	0
station 5	0	0	0	0	0	0

Response time	Tract 1	Tract 2	Tract 3	Tract 4	Tract 5
station 1	5	12	30	20	15
station 2	20	4	15	10	25
station 3	15	20	6	15	12
station 4	25	15	25	4	10
station 5	10	25	15	12	15
Fire Frequency	2	1	3	1	3
min time	5	12	6	15	12

Objective function 91

Constraints

2	=	2
1	=	1
1	=	1
1	=	1
1	=	1
1	=	1
1	<=	1
1	<=	1
0	<=	1
0	<=	1
0	<=	1
0	<=	0
0	<=	0
0	<=	0
0	<=	0
0	<=	0
0	<=	1
0	<=	1
1	<=	1
1	<=	1
1	<=	1
0	<=	0
0	<=	0
0	<=	0
0	<=	0
0	<=	0
0	<=	0
0	<=	0
0	<=	0
0	<=	0

Figure 6-23: Excel solver solution (2).

If we use the P- Median Algorithm we follow the algorithm [25] :

Multimedian Heuristic Algorithm

STEP 1: Let m = 1. Find the l-median of the network G(N, A) using Single Median Algorithm. Let the l-median be at node i. Set S = {i}.

STEP 2: Add a new facility to the current membership of the set S by choosing that location among the nodes in N – S, the nodes which are not in S, which produces the maximum possible improvement in the objective function as the number of medians increases by 1. Let m = m + 1.

STEP 3: Attempt to improve the objective function by substituting in a systematic way, *one* at a time, one of the nodes in S with a node that is in N-S. Every time an improved solution is obtained, use this as the new "incumbent" solution, S, and repeat Step 3. When all possible single-node substitutions for a set S have been attempted without improving the objective function, go to Step 4.

STEP 4: If m = k, stop; otherwise, return to Step 2.

Single Median Algorithm

STEP 1: Obtain the minimum distance matrix for the nodes of G.

STEP 2: Multiply the jth column of the minimum distance matrix by the demand weight h_j (j = 1, 2, . . ., n) to obtain the matrix $h_j \cdot d(i\ j)]$.

STEP 3: For each row i of the $[h_j \cdot d(i\ j)]$ matrix, compute the sum of all the terms in the row. The node that corresponds to the row with the minimum sum of terms is the location for the l-median and we have the following solution:

	2	1	3	1	3
	A	B	C	D	E
A	5	12	30	20	15
B	20	4	15	10	25
C	15	20	6	15	12
D	25	15	25	4	10
E	10	25	15	12	15

$\Rightarrow$

	A	B	C	D	E
A	10(= 2 * 5)	12	90	20	45
B	40	4	45	10	75
C	30	20	18	15	36
D	50	15	75	4	30
E	20	25	45	12	45

$\Rightarrow$

						Sum
A	10	12	90	20	45	177(= 10 + 12 + 90 + 20 + 45)
B	40	4	45	10	75	174
C	30	20	18	15	36	119
D	50	15	75	4	30	174
E	20	25	45	12	45	147

We set S = {C}

In the next step we estimate all possible combinations for n = 5 and k =2 such as:

$$\binom{5}{2} = \frac{5!}{(5-2)!2!} = 10 \text{ combinations}$$

A – B: 10+4+45+10+45 = 114

A – C: 10+12+18+15+36 = 91

A – D: 10+12+75+4+30 = 131

A – E: 10+12+45+12+45 = 124

B – C: 30+4+18+10+36 = 98

B – D: 40+4+45+4+30 = 123

B – E: 20+4+45+10+45 = 124

C – D: 30+15+18+4+30 = 97

C – E: 20+20+18+12+36 = 106

D – E: 20+15+45+4+30 = 114

The value that produces the maximum possible improvement in the objective function is 91. This means that the two fire stations will be built in the tracts 1 and 3. Fire station in tract 1 will respond to fires in tracts 1, 2 and fire station located in tract 3 will respond to fires in tract 3, 4 and 5. The results are illustrated in the following Table 6-10.

Fire Station in:		**Will respond to fires in Tracts:**				
		1	**2**	**3**	**4**	**5**
Track 1	**Open**	1	1	0	0	0
Track 2	Close	0	0	0	0	0
Track 3	**Open**	0	0	1	1	1
Track 4	Close	0	0	0	0	0
Track 5	Close	0	0	0	0	0

Table 6-10: Computational Results.

6.3 An application of P-Median and Hypercube problem

Now we will try to solve the problem, described in section 6.2, for three servers using the P-median model and then we will use the hypercube model, described in section 6.1, in order to compare our results.

Suppose that we are to allocate three fire vehicles in the city. The city has been divided into five tracts as previously, with no more than one fire station to be located in any given tract. Each station is to respond to all of the fires that occur in the tract in which it is located, as well as in the other tracts that are assigned to the station. Table 6-9 gives the same average response time to a fire in each tract if their tract is served by a station in a given tract while the bottom row gives the same forecasted average number of fires that will occur in each of tracts each day. The objective is to minimize the overall average of the response times to fires.

We define :

s, t: indices for tract T = {1,2,...5}

d(s, t): response time/ distance between s and t

f(t): frequency of fire in tract t.

$$x(s,t) = \begin{cases} 1 & \text{Fire in tract 't' will be covered by station in tract 's'} \\ 0 & \text{otherwise} \end{cases}$$

$$y(s) = \begin{cases} 1 & \text{Build a fire station in tract 's'} \\ 0 & \text{otherwise} \end{cases}$$

and the model now can be formulated as follows:

Minimize $Z = \sum_{s \in T} \sum_{t \in T} x(s,t) \cdot d(s,t) \cdot f(t)$

Subject to:

$$\sum_{s \in T} y_s = 3$$

$$\sum_{s \in T} x(s,t) = 1 \qquad \forall t \in T$$

$$x(s,t) \leq y_s \qquad \forall s,t \in T$$

$$x(s,t),\ y(t) \in \{0,1\} \quad \forall s,t \in T$$

The objective function minimizes the overall average response time to fires. The first constraint states that we are going to locate exactly 3 fire stations. The second constraint states that each tract has to be covered by a fire station so that some one will respond to the fire. The third constraint states that demand tracts can only be assigned to open fire stations and the forth constraint states that no more than one station can be allocated in each tract.

This time, we solved the problem using Excel Solver and the solution is depicted in Figure 6-24. The value that produces the maximum possible improvement in the objective function is 74. This means that the three fire stations will be built in tracts 1, 3 and 4. Fire station in tract 1 will respond to fires in tracts 1, 2, fire station located in tract 3 will respond to fires in tract 3 and fire station located in tract 4 will respond to fires in tracts 4 and 5. The results are illustrated in the following Table 6-11.

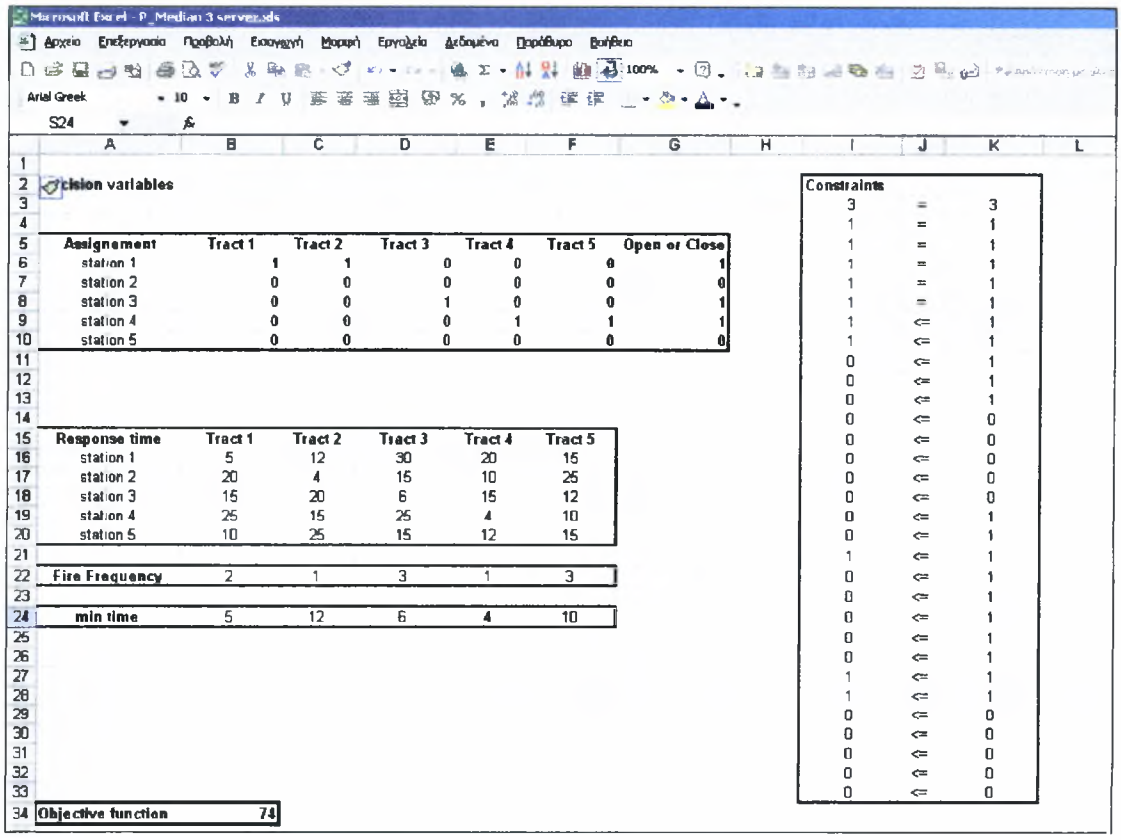


Figure 6-24: Excel solver solution for three servers.

Our example described above, represents a three-server model of a server-to- customer spatially distributed queueing system. This model is generalized to N servers with the hypercube model discussed in the previous section.

Fire Station in:		**Will respond to fires in Tracts:**				
		1	**2**	**3**	**4**	**5**
Track 1	**Open**	1	1	0	0	0
Track 2	Close	0	0	0	0	0
Track 3	**Open**	0	0	1	0	0
Track 4	**Open**	0	0	0	1	1
Track 5	Close	0	0	0	0	0

Table 6-11: Computational Results.

Now we will implement the hypercube model, using the assignment resulted from P-Median problem for three servers where server 1 will respond to fires in tracts 1, 2, server 2 will respond to fires in tract 3 and server 3 will respond to fires in tracts 4 and 5. We will examine, by this model, if the workloads of these three servers are evenly distributed.

For convenience, we change frequency of fires values proportionally and now instead of f(1) = 2, f(2) = 1, f(3) = 3, f(4) = 1, f(5) = 3, we will have the following values: $\lambda_1 = 0,3$, $\lambda_2 = 0,15$, $\lambda_3 = 0,45$, $\lambda_4 = 0,15$, $\lambda_5 = 0,45$, and $\lambda = 1,5$, Figure 6-25. The dispatch preferences for three-server city are shown in Table 6-12 and results from Table 6-9.

Tract Number	**First Preference Unit**	**Second Preference Unit**	**Third Preference Unit**
1	1	2	3
2	1	3	2
3	2	3	1
4	3	2	1
5	3	2	1

Table 6-12: Dispatch preferences for three-server city.

In this application where N = 3 unit problem, we suppose that the following events occur:

1. A request for service arrives from tract 4.
2. A request for service arrives from tract 2.
3. Server 3 completes service on its request.
4. A request for service arrives from tract 1.
5. A request for service arrives from tract 5.

and the sequence of states occupied by the system is as follows:

1. State where no unit is busy ⟶ (0,0,0)
2. A request for service arrives from tract 4 ⟶ (1,0,0)
3. A request for service arrives from tract 2 ⟶ (1,0,1)
4. Server 3 completes service on its request ⟶ (0,0,1)
5. A request for service arrives from tract 1 ⟶ (0,1,1)
6. A request for service arrives from tract 5 ⟶ (1,1,1)

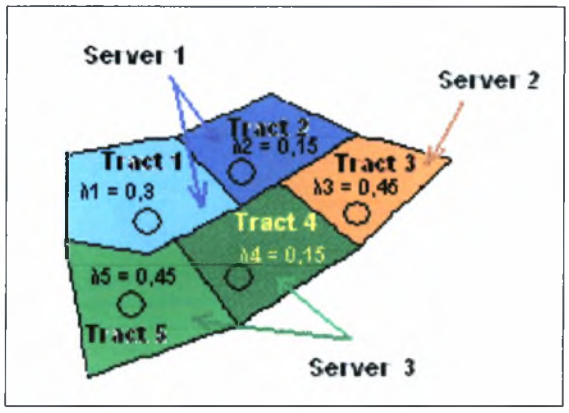


Figure 6-25: Map of three server city.

We compute the values of the transition rates as previously, assuming that μ=1 and the state-transition diagram is shown in Figure 6-26 where we define:

$r_1 = \lambda_4+\lambda_5 = 0,60$

$r_2 = \lambda_1+\lambda_2 = 0,45$

$r_3 = \lambda_3 = 0,45$

$r_4 = \lambda_3 = 0,45$

$r_5 = \lambda_1 + \lambda_3 = 0,75$

$r_6 = \lambda_1+\lambda_2 = 0,45$

$r_7 = r_6 + \lambda_4 = 0,60$

$r_8 = r_9 + \lambda_1 + \lambda_2 = 1,05$

$r_9 = \lambda_4 + \lambda_5 = 0,60$

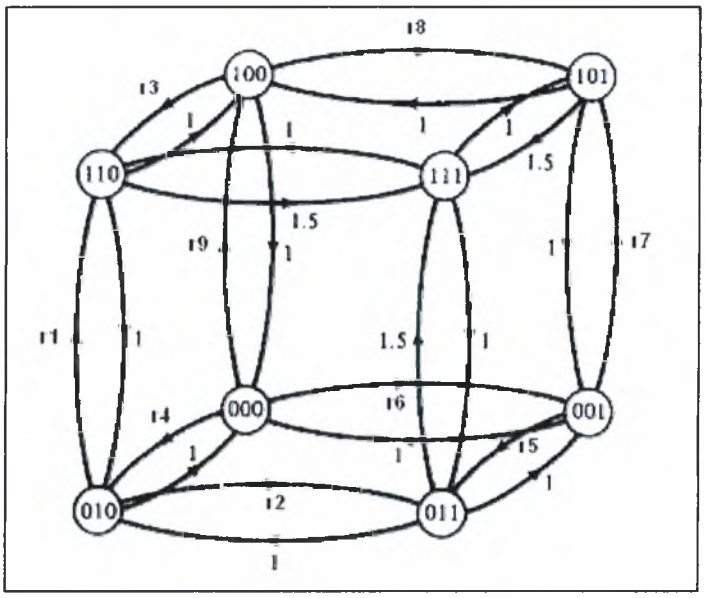
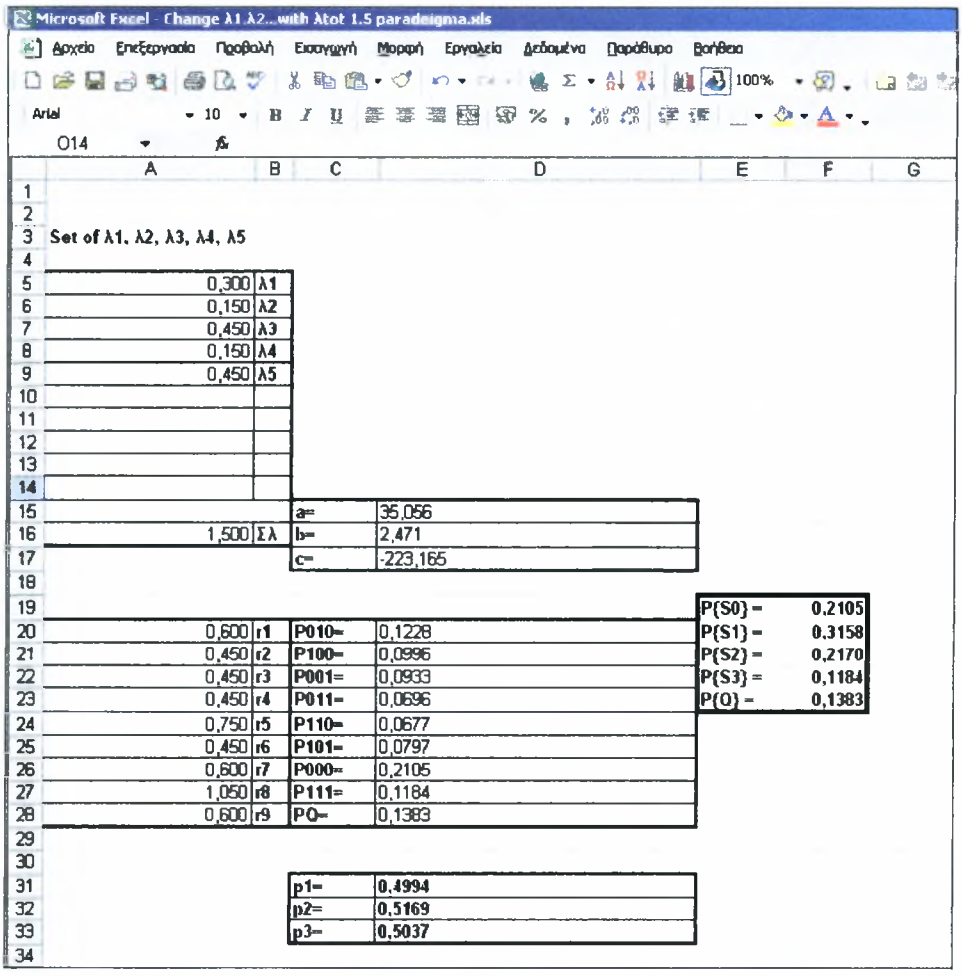


Figure 6-26: Three server state-transition diagram.

Solving the hypercube equilibrium equations for nonsaturated system states using the Excel spreadsheet we arrive at the following values for the state probabilities depicted in Figures 6-27 and 6-28.

As we can see, the values of the state probabilities are: $P_{000} = 0,2105$, $P_{001} = 0,0933$, $P_{010} = 0,1228$, $P_{100} = 0,2017$, $P_{110} = 0,0677$, $P_{101} = 0,0797$, $P_{011} = 0,0696$, $P_{111} = 0,1184$, $P_Q = 0,1383$. Furthermore, we notice that the workloads of servers 1, 2, 3 are $\rho_1 = 0,4994$, $\rho_2 = 0,5169$, $\rho_3 = 0,5037$, respectively. We notice that the workload sharing among response units caused the workloads of the units to be more evenly distributed than the workloads of the primary response areas. If each unit served only the customers of its own response area, the workloads would have been $\rho_1 = 0,45$, $\rho_2 = 0,45$, $\rho_3 = 0,60$.



Microsoft Excel - Change λ1,λ2...with λtot 1.5 paradeigma.xls

Set of λ1, λ2, λ3, λ4, λ5

0,300	λ1	
0,150	λ2	
0,450	λ3	
0,150	λ4	
0,450	λ5	
1,500	Σλ	

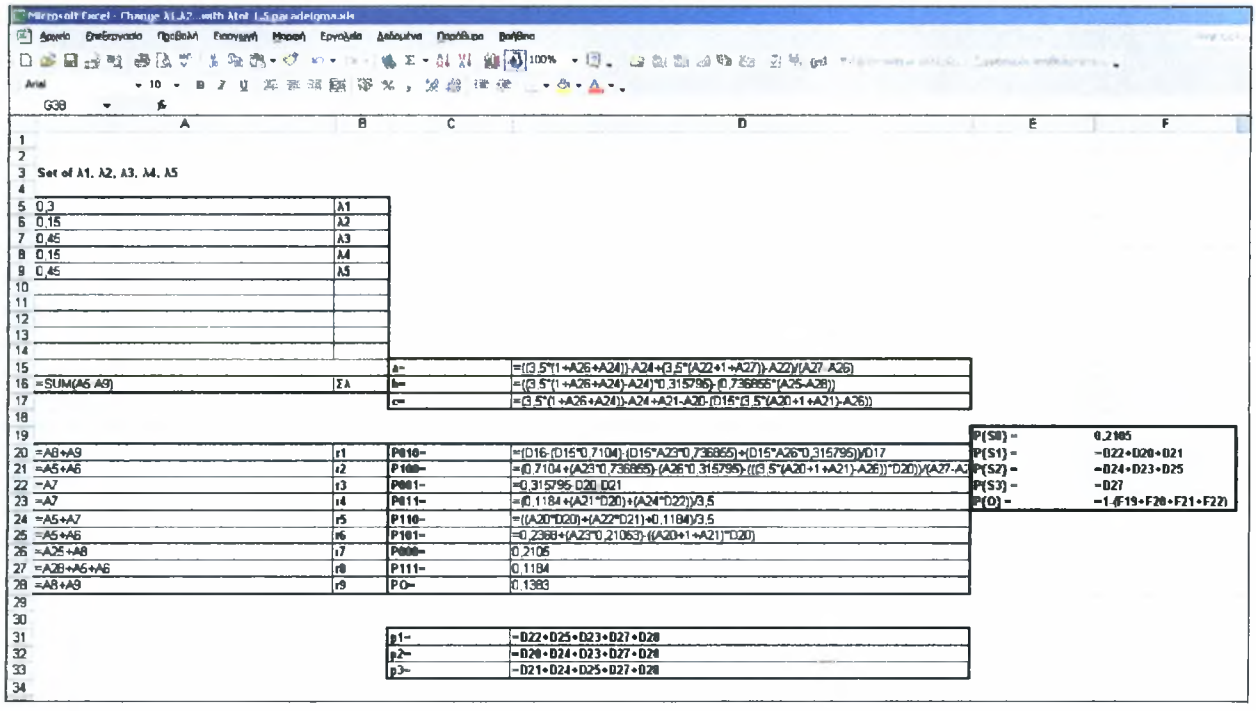
a=	35,056
b=	2,471
c=	-223,165

0,600	r1	P010=	0,1228
0,450	r2	P100=	0,0996
0,450	r3	P001=	0,0933
0,450	r4	P011=	0,0696
0,750	r5	P110=	0,0677
0,450	r6	P101=	0,0797
0,600	r7	P000=	0,2105
1,050	r8	P111=	0,1184
0,600	r9	PO=	0,1383

p1=	0,4994
p2=	0,5169
p3=	0,5037

P{S0} =	0,2105
P{S1} =	0,3158
P{S2} =	0,2170
P{S3} =	0,1184
P(O) =	0,1383

Figure 6-27: Estimation of state probabilities and servers' workloads.



Set of λ1, λ2, λ3, λ4, λ5

0,3	λ1	
0,15	λ2	
0,45	λ3	
0,15	λ4	
0,45	λ5	
=SUM(A5:A9)	Σλ	

a=	=((3,5*(1+A26+A24))-A24+(3,5*(A22+1+A27))-A22)/(A27-A26)
b=	=((3,5*(1+A26+A24)-A24)*0,315795)-(0,736855*(A25-A28))
c=	=(3,5*(1+A26+A24))-A24+A21-A20-(D15*(3,5*(A20+1+A21)-A26))

=A8+A9	r1	P010=	=(D16-(D15*0,7104)-(D15*A23*0,736855)+(D15*A26*0,315795))/D17
=A5+A6	r2	P100=	=(0,7104+(A23*0,736855)-(A26*0,315795)-((3,5*(A20+1+A21)-A26))*D20))/(A27-A2
=A7	r3	P001=	=0,315795-D20-D21
=A7	r4	P011=	=(0,1184+(A21*D20)+(A24*D22))/3,5
=A5+A7	r5	P110=	=((A20*D20)+(A22*D21)+0,1184)/3,5
=A5+A6	r6	P101=	=0,2368+(A23*0,21063)-((A20+1+A21)*D20)
=A25+A8	r7	P000=	0,2105
=A28+A5+A6	r8	P111=	0,1184
=A8+A9	r9	PO=	0,1383

p1=	=D22+D25+D23+D27+D28
p2=	=D20+D24+D23+D27+D28
p3=	=D21+D24+D25+D27+D28

P{S0} =	0,2105
P{S1} =	=D22+D20+D21
P{S2} =	=D24+D23+D25
P{S3} =	=D27
P(O) =	=1-(F19+F20+F21+F22)

Figure 6-28: Formulas of spreadsheet of Figure 6-27.

Chapter 7 Conclusions

Location of emergency service is an important aspect in the life of every city and due to limited resources, requires a careful management.

Optimum locations differ as to whether or not decision-makers want to minimize average response time, maximum response time, average distance, or maximum distance, or want to locate in areas where more accidents occur, want to minimize station overlaps.

Mathematical models may be useful in dealing with such a problem. There has been an important evolution in the development of emergency vehicles location and relocation models over the past years. In this thesis, we attempted to provide an overview of emergency vehicles location models dedicated to capturing the complex time and uncertainty characteristics of most real-world problem instances. Furthermore, we examined two applications of emergency response vehicles location models HQM and P-Median for a profound comprehension. In the end, we made a combination of previous models to compare our results.

The first models were very basic and did not take into account the fact that some coverage is lost when an emergency vehicle is dispatched to a call. Nevertheless, these early models served as a basis for the development of all subsequent models. The question of emergency vehicles non-availability was addressed in two main ways. Deterministic models yield solutions in which demand

points are overcovered, but the actual availability of emergency vehicles is not considered. Dynamic models have just started to emerge. They can be used to periodically update emergency vehicles positions throughout the day. Probabilistic models work with the busy fraction of vehicles, which can be estimated in a number of ways. Development of tractable models which consider both the stochastic and dynamic aspects of emergency vehicle location would be a long term plan.

One of the most important indicators of the performance of any emergency service system is response time. Since response time can have a significant impact on the loss of life and property at an emergency, it is used as a principal measure of effectiveness in many models developed for analyzing the deployment of emergency vehicles. In this thesis we also, focused on the description of some approaches to estimate travel distance and travel time.

A very important issue in modeling is data collection. If you cannot obtain data to run in the model, then there is no need for the model. Little work has been done on long term demand forecasting. Most models use deterministic data or the average of a sample since there are few good estimating procedures to obtain distributions. Accurate travel and service time estimates are critical for building valid detailed models, but little work has been done in this area.

As we mentioned above, we examined two applications of emergency response vehicles location models. The first one, was an application of hypercube queueing model. We considered a three-server city partitioned into 10 geographical tracts, each having a rate λ_i, of arrival of requests from each tract i. We implemented the hypercube queueing model and in association with an Excel spreadsheet we concluded to obtain any time for random values of λ_i and constant λ_{tot}, the state probabilities of the system and the workloads of the individual servers. We continued

with a generalized case of the previous example where different values of λ_{tot} and λ_1, λ_2.....λ_{10} result to different state probabilities of the system, workloads of the three servers and the probability that a queue of positive length exists.

The second application was representative of P-median model. We supposed a city partitioned into five tracts where two fire vehicles were to be located. The problem was solved using Excel Solver and P-Median Algorithm.

In the end, we solved P-Median model with Excel Solver tool for three servers and we implemented the hypercube model using the assignment resulted from P-Median problem. We concluded that, the workloads of the three servers are evenly distributed which confirms that the assignment of P-Median problem is sufficient.

During implementation time I contacted Fire Brigade chiefs in Volos and Larissa, biggest cities of Thessaly region. We arrange meetings in their office and we discussed the way they now allocate vehicles in different areas (industrial zones, mountains) during the various season and time periods and the way they make zoning of their responsibility area.

Findings were quite disappointing. Zoning and Selection of places is not decided as the optimum solution of a set of important parameters as we discuss in our work but is done empirically. Sometimes factors such as the communication restrictions influence vehicles positioning so deep that no optimization in terms of minimizing distance and time while covering maximum area susceptible to fire, can be considered.

Situation was even worse in terms of data records that were not kept for the incidents that fire brigade intervene, beside the book of incidents were they just write what kind of event it was and the investigation report if there is need for an

investigation. Critical data for optimum positioning, such as time of access, time needed to be informed, time of vehicles occupancy and so forth are not kept.

That is the main reason that Fire Brigade has major difficulties to decide a different positioning allocation of its vehicles and why selection of positions for building new fire brigade stations is not being done in an operationally meaningful framework as we can see from the open procedure for selecting the place for the new Main Station of Volos City Fire Brigade Station.

Interviews with Fire Brigade Chiefs and head officers lead to conclusion that present work could be extremely valuable for their work especially in cases where zoning of an area (urban or rural) changes, as it happens during events (e.g. athletic games, festivals) or operation of an industry with dangerous materials, or changes in transport infrastructures (e.g. tunnels, bridges), new hotels in mountains etc.

It is clear that in time of economic scarcity and restricted resources use of tools and methodologies such as these presented in our work are useful and will support decision making of Fire Brigade both in Planning and Emergency Management fields. At the moment we still lack central guidance from the central headquarters

Finally, we present certain tasks that should be done to improve positioning of emergency vehicles, especially in the case of Fire Brigade, so that our models and tools as well as others of similar nature could be used to optimize social benefit.

1. Cities zoning could be done after continuous risk assessment following certain criteria, such as - indicatively - population density, critical infrastructures, ease of access, land uses etc.

2. Data of Fire brigade missions should be kept in detail in electronic forms that will make easier use of "smart" tools for allocation and positioning of resources or even patrols design and redesign depending on the conditions.

3. Past events/ records investigation by researchers to identify and measure if possible parameters that play critical role in vehicles positioning and effectiveness in event management.

4. Fire brigade should test such tools in real applications especially in summer times where those tools can improve the proximity time to fires and effectiveness of the fire prevention and management system.

5. There is need for public presentation of a case study (e.g. Pelion mountain) using real data with past events (fires), access times, etc and then show how presented tools could improve the situation both in terms of maximum area coverage, equal personnel workloads, minimum intervention times and other critical parameters so that Fire Brigade being "forced" to cooperate in the direction of making their work more effective and efficient.

6. Use of tools for vehicle allocation can be used by Fire Brigade to evaluate their resources, and make a better planning. Request to State for vehicles, men, new positioning of headquarters etc can be supported by scientific evidence based on real data. Such approach will maximize effectiveness and minimize public funds spending - in different ways - in emergencies management.

References

[1] Aggarwal, S. C. (1982). "A focussed review of scheduling in services" European Journal of Operational Research, Vol. 9, pp. 114-121.

[2] Berlin, G., Revelle, C. and Elzinga, J. (1976). "Determining ambulance-hospital locations for on-scene and hospital services". Environment and Planning A, 8, pp. 553-561.

[3] Berman, O. , Larson, R.C and Chiu, S.S (1985). "Optimal server location on a network operating as an M/G/1 queue". Operations Research, Vol. 33, pp. 746-771.

[4] Brotcorne, L., Laporte, G. and Semet, F. (2003). "Ambulance location and relocation models". European Journal of Operational Research 147, pp. 451-463.

[5] Carson, Y. and Batta, R. (1990). "Locating an ambulance on the Amherst campus of the State University of New York at Buffalo". Interfaces, 20, pp. 43-49.

[6] Church, R. and Revelle, C. (1974). "The maximal covering locational problem". Papers of the Regional Science Association, Vol 32, pp. 101-108.

[7] Current, J., Daskin M. and Schilling D. (2002). "Discrete Network Location Models". Facility location: Applications and Theory. Edited by Z. Drezner and H. W. Hamacher.

[8] Daskin, M. S and Dean, L. K. (2004). "Location of Health Care Facilities". Chapter 3 in the Handbook of OR/MS in Health Care : A Handbook of Methodi and Applications, F. Sainfort, M. Brandeau and W. Pierskalla, editors, Klumer, pp. 43-76.

[9] Daskin, M.S. and Stern, EH (1981). "A hierarchical objective set covering model for emergency medical service vehicle deployment". Transportation Science, Vol 15., No 2, pp. 137-152.

[10] Davis G. S. (1981). "Analysis of the Deployment of Emergency Medical Services". Omega, Vol. 9, No 6.

[11] Eaton, D. , Daskin, M. S, Simmons, D. Bulloch, B. and Jansma, G. (1985). "Determining emergency medical deployment in Austin, Texas". Interfaces, 15(1), pp. 96-108.

[12] Galvão, R.D., Chiyoshi, F. Y., Morabito R. (2005). "Towards unified formulations and extensions of two classical probabilistic location models". Computers & Operations Research, 32, pp. 15-33.

[13] Geroliminis N, Karlaftis M, Skabardonis A. (2006). "A Generalized Hypercube Queueing Model for locating emergency response vehicles in Urban Transportation Networks". Presentation in 85^{n} Annual Meeting Transportation Research Board, Washington, D. C, January 2006.

[14] Gendreau, M. , G. Laport and F. Semet (2001). "A dynamic model and parallel tabu search heuristic for real-time ambulance relocation". Parallel Computing, Vol. 27, pp. 1641-1653.

[15] Garfinkel, RS, Neebe, A.W. and Rao, M.R. (1977). "The m-center problem: Minimax facility location ". Management Science, 23, pp. 1133-1142.

[16] Hakimi L. S. (1964). "Optimum Locations of Switching Centers and the Absolute Centers and Medians of a Graph". Operations Research, Vol. 12, No 3, pp. 450-459.

[17] Hillsman E. and Rhoda, R. (1978). "Errors in measuring distances from population to service centers". Annals of Regional Science vol. 12 pp. 74-88.

[18] Hochbaum, D.S. and Pathria, A. (1998). "Locating centers in a dynamically changing network and related problems". Location Science, 6 , pp. 243-256.

[19] Hogan, K. and Revelle, C (1986). "Concepts and applications of backup coverage". Management Science, 32, pp. 1434-1444.

[20] Klose, A. and Drexl, A. (2004). "Facility location models for distribution system design". European Journal of Operational Research, 162 , pp. 4-29.

[21] Kolesar, P. and Walker, W. E. (1974). "An algorithm for the dynamic relocation of fire companies". Operations Research, Vol. 22, pp. 249-274.

[22] Kolesar, P. and Blum, H. E. (1973). "Square root laws for fire engine response distances". Management Science, Vol. 19, pp. 1368-1378.

[23] Larson, R.C (1974). "A hypercube queuing model for facility location and redistricting in urban emergency services". Computers and Operatios Research, 1 , pp. 67-95.

[24] Larson, R. C (1975). "Approximating the performance of urban emergency service systems". Operations Research, Vol. 23, No 5, pp. 845-868.

[25] Larson, R.C and Odoni, A. R. (1981). "Urban Operation Research". http://web.mit.edu/urban_or_book/www/book/.

[26] Marianov, V and Revelle, C. (1996). "The queueing maximal availability location problem : A model for the siting of emergency vehicles". European Journal of Operational Research, 93, pp.110-120.

[27] Marianov, V. and Serra, D. (2002). "Location Problems in the Public Sector". Facility location : Applications and Theory. Edited by Z. Drezner and H. W. Hamacher.

[28] Marsh, M. and Schilling, D. (1994). "Equity Measurement in facility location Analysis : A review and Framework". European Journal of Operational Research Vol. 74-1, pp. 1-17.

[29] Mirchandani, P. B. (1980). "Locating decisions on stochastic networks". Geographical Analysis, 12, pp. 172-183.

[30] Owen, H. S. and Daskin, S. M. (1998). "Strategic facility location: A review". European Journal of Operational Research, 111, pp. 423-447.

[31] Othman, I. A and Rand, K. G. (2003). "A goal programming model applied to the EMS system at Riyadh City, Saudi Arabia". Lancaster University Management School, Workinh Paper 2003.

[32] Repede, F. J. and Bernardo, J. J. (1994). "Developing and validating a decision support system for locating emergency medical vehicles in Lousville, Kentucky". European Journal of Operational Research 75, pp. 567-581.

[33] Revelle, C. 1989. "Review, extension and prediction in emergency service siting models". European Journal of Operational Research Vol. 40-1, pp. 58-69.

[34] Revelle, C. and Hogan, K. (1986). "A reliability constrained siting model with local estimates of busy fractions". Environment and Planning, B15, pp. 143-152 (abstract).

[35] Revelle, C. and Hogan, K. (1989). "The maximum reliability location problem and a-reliable p-center problem : Derivatives of the probabilistic location set covering problem". Annals of Operations Research, 18, pp. 155-174.

[36] Revelle, C. (1993). "Facility siting and integer-friendly programming". European Journal of Operational Research, vol. 65, pp. 147-158.

[37] Schilling, D.A. , Elinga, D.J. , Cohon, J. , Church, R.L., Revelle, C.S., (1979). "The TEAM/FLEET models for simultaneous facility and equipment sitting. Transportation Science, 13, pp. 163-175.

[38] Serra, D. and Marianov, V. (1996). "The P-Median Problem in a Changing Network: The Case of Barcelona"

http://www.econ.upf.edu/docs/papers/downloads/180.pdf

[39] Toregas, C. Swain, R., Revelle, C. and Berman, L. (1971). "The location of emergency service facilities". Operation Research, Vol 19-2, pp. 1363-1373.