Lancaster University Management School
Working Paper
2003/034

Extensions to emergency vehicle location models

Othman Ibraheem Alsalloum and Graham K Rand

The Department of Management Science
Lancaster University Management School
Lancaster LA1 4YX
UK

Extensions to emergency vehicle location models

Othman Ibraheem Alsalloum (King Saud University, Riyadh, Saudi Arabia)
and Graham K. Rand (Lancaster University, UK)

Abstract

This paper is concerned with extending models for the Maximal Covering Location Problem in two ways. First, the usual 0-1 coverage definition is replaced by the probability of covering a demand within the target time. Second, once the locations are determined, the minimum number of vehicles at each location that satisfies the required performance levels is determined. Thus, the problem of identifying the optimal locations of a pre-specified number of emergency medical service stations is addressed by goal programming. The first goal is to locate these stations so the maximum expected demand can be reached within a pre-specified target time. Then, the second goal is to ensure that any demand arising located within the service area of the station will find at least one vehicle, such as an ambulance, available. Erlang's loss formula is used to identify the arrival rates when it is necessary to add an ambulance in order to maintain the performance level for the availability of ambulances. The model developed has been used to evaluate locations for the Saudi Arabian Red Crescent Society, Riyadh City, Saudi Arabia.

Keywords:Location, Emergency Medical Services

1. Introduction

This paper describes a model that has been developed and applied to the EMS of Riyadh, the capital city of Saudi Arabia. The application is described more fully in Alsalloum and Rand (2003). The aim of the emergency medical service (EMS) of Saudi Arabia is to reduce mortality and health deterioration caused by emergency incidents or illness. This goal can be achieved if suitable care arrives on time at the location of the incidents. Therefore, rapid response to an incident is one important measurement of an EMS system success. However, an EMS is provided within a tight public sector budget. Therefore, a rational and optimal way of locating EMS stations and allocating EMS ambulances to these stations is required.

The model developed here is an extension to the Maximal Covering Location Problem (MCLP), which was presented by Church and Revelle (1974). The purpose is to identify the optimal locations of a specified number of EMS stations. The first goal is to locate these stations so the maximum expected demand may be reached within a pre-specified target time. The traditional definition used in the set covering problem models is that the demand node is covered if it is within the target time or distance, otherwise it will not be covered. Here, the usual 0-1 coverage definition is replaced by the probability of covering a demand within the target time. The second goal is to ensure that any demand arising located within the target time will find at least one ambulance available. Erlang's loss formula is used to identify the demand (i.e. the arrival rates) which makes it necessary to add an ambulance in order to maintain the required performance level for the availability of ambulances. The problem is formulated as a goal programming problem to optimise the locations, and then to find the minimum number of vehicles satisfying the performance levels.

2. Emergency Medical Service models

EMS models typically fall into two categories: deterministic or stochastic. Deterministic mathematical programming models are attractive since they recommend a "best" decision given a set of constraints and quantifiable performance measures. However, their weakness lies in the fact that they often fail to take into consideration the probabilistic nature of the EMS environment. In contrast, stochastic models better address this issue of the probabilistic nature of the EMS environment. They take into account the probabilities of servers being busy, and/or the stochastic nature of ambulances arriving to the demand points. Each of the above categories falls into two subcategories: either to find the number of servers required so as to cover the whole region, or to optimise the number of servers available to serve the greatest demand. Marianov and ReVelle (1995) present an excellent survey of these models. Here a brief background to some of these models is given in order to motivate the developments of the model described in this paper.

The *p*-median model, first introduced by Hakimi (1964), looks for a set of *p* points that yield the smallest possible weighted distance: the optimal *p*-facility solution set. The *p*-median solution finds the locations in such a way that the total travel time from all demand areas to these locations is minimised. Toregas and ReVelle (1972) reduced the complexity associated with the *p*-median problem by introducing a new model for the Location Set Covering Problem (LSCP). The LSCP imposes a maximum distance (or time or cost) on a *p*-median problem so as to include only those demand nodes within the maximum pre-specified distance. The solution to the LSCP obtains the minimum number of facilities to be opened to 'cover' all the area within a pre-specified distance.

Although developments were made in these *p*-median and LSCP models, some important issues were ignored. First, what if the available resources (facilities) are less than the minimum required? Second, will locating only one facility within a neighbourhood be enough? To rectify these weaknesses, researchers investigated two issues in particular: covering all nodes within a specific distance and ensuring the availability of an ambulance when a call is received. The requirement to cover all nodes or demands (as the LSCP states) within a pre-specified limited distance or time is often costly. Furthermore, some of these required facilities will be used to cover only a few nodes that may have very small demands. Therefore, Church and ReVelle (1974) developed a model for the Maximal Covering Location Problem (MCLP). The MCLP model finds the location of a pre-specified number of facilities *n* so as to maximise the demand covered by at least one facility. Since the available resources are often less than the required number of facilities to cover all demand, *n* is equal to or less than the minimum number of facilities required by the previous LSCP model.

All MCLP algorithms assume that vehicles located at a base will be available to serve a call from zones they have been assigned to cover, and will never be busy. However, this may not be the case in practice, and the most desirable ambulance to dispatch to a call in zone *i* may be busy when a call from zone *i* is received. Therefore, the probability of a server being busy should be considered. When an emergency call in a region occurs while the designated ambulance is engaged in service, locating a single ambulance within a specific time or distance will not be enough, and it is necessary to have at least one ambulance available with some probability within the time or distance standard.

To ensure the availability of an ambulance when a new call is received, extensions of the LSCP model have been developed. A hierarchical objective set covering model (HOSC) by Daskin and Stern (1981) and multiple coverage or backup, as it is called by Hogan and ReVelle (1986), have been created. Backup coverage is used as a basis by which coverage may be protected from varying demand intensities during different times. By multiple coverage one can increase the probability of the presence of at least one vehicle within the distance or time standard, even when there is congestion. Narasimhan et al. (1992), extended single service to multiple levels of backup services. In addition, the uncovered demands are forced to be assigned to a facility even if they are beyond the pre-specified target time. Since the problem is NP hard, a Lagrangian relaxation approach is used to develop a heuristic solution procedure. This approach solves the problem effectively. However, these models do not determine the number of ambulances to be placed at any open base. Nor do these levels of coverage take into account that population or call frequency varies from one demand node to another.

The Maximum Expected Covering Location Problem (MEXCLP) developed by Daskin (1983), seeks to maximise the expected value of coverage within a time standard, using a heuristic approach. Daskin assumed that the busy probability is the same for all servers in the system. ReVelle and Hogan (1988, 1989) extended the notion of MEXCLP by introducing the probabilistic location set covering problem (PLSCP) model to utilise a region specific busy fraction instead of a system wide busy fraction. PLSCP is similar to the LSCP model, but includes a set of constraints on the reliability of a server being available. Since PLSCP will usually lead to a potentially large number of servers being assigned or required, ReVelle and Hogan extended PLSCP to a more realistic model. The Maximum Availability Location Problem (MALP) model seeks to locate servers in such a way as to maximise the population covered with a stated reliability. The difference between MEXCLP, PLSCP, and MALP lies in the way they include the busy fraction in the formulation. MEXCLP includes the probabilities in the objective function, while PLSCP and MALP include the busy fraction in the constraints. In addition, MEXCLP uses the busy fraction to maximise the expected demands covered, while PLSCP and MALP use it to meet the reliability constraints. However, PLSCP and MALP do not assign demand nodes to open centres, so they assume that any node within the target time will be included in the total demand rate. In other words, some demand nodes are counted more than once, especially if these nodes are within the target time from more than one opened centre.

Ball and Lin (1993) formulated a new version of PLSCP, in which an upper bound of the "uncovered probability" of each demand is constrained to be less than an upper bound value. This model is called Rel-P and it is an extension of PLSCP in two ways. First, it assigns demands to the opened centres. Second, because it assigns each demand to an open centre, it takes into consideration the exact demands assigned to any open station. This helps when deciding the minimum number of facilities located in any opened station. Various proposed pre-processing techniques reduced the computation time required by branch and bound solution algorithms. As a result, the branch and bound codes used solved the problem efficiently.

Marianov and ReVelle (1994) relaxed the assumption in the PLSCP of independence between the probabilities of different servers being busy. They modelled the behaviour in each region as an M/M/S-loss queueing system. The use of an acceptable probabilistic structure inside an optimisation model for facility siting is the distinctive contribution of this model. This model is called the queueing probabilistic location set covering problem (Q-PLSCP). In addition, to find the maximum availability level α, which gives the desired number of servers when applied to Q-PLSCP, a procedure, MASH, is devised, which maximises the minimum system-wide reliability level obtainable with the desired number of servers. Marianov and ReVelle (1996) further developed the MALP. Their new model is called Queueing Maximal Availability Location Problem (Q-MALP). The main difference between MALP and Q-MALP resides in the methodology for the calculation of the smallest integer that satisfies the required reliability. In addition, in this model they treated the distances/times as random. The smallest integer satisfying the required reliability is calculated using the M/G/S-loss queueing system. Therefore, the independence assumption for servers' busy fractions in the original MALP model is avoided in Q-MALP.

Here a realistic and a practicable model is developed, which takes into account not only the probabilistic nature of the problem, but also the fact that the available resources are often limited. The model is an extension to the MCLP models that locate the EMS facilities and utilises the work of Charnes and Storbeck (1980), Ball and Lin (1993), and Marianov and ReVelle (1996) to allocate the exact numbers of ambulances required in each open base. The MCLP models were developed as a result of unrealistic assumptions associated with the LSCP models, since the LSCP models ignore the case when the numbers of available resources or facilities are less than the minimum number required.

A review of real-world applications can be found in Alsalloum and Rand (2003).

3. Defining coverage

In set covering location problems, demands that need to be covered are often grouped in areas due to the impossibility of dealing with each single demand separately. The aggregated demands of each area are usually located at the centre of the area. So, when trying to determine the demands covered within the target distance, the distances to and from the centres of areas that represent these demands are used. While this approach is necessary to allow the problem to be solved, there are some potential disadvantages.

The most important issue in this context is the way that the coverage is defined by the traditional set covering location models. The traditional definition used in the set covering problem models is that the demand node is covered if it is within the target time or distance, otherwise it will not be covered. In other words, the probability of covering a demand node within the target distance is 100%, and the probability of covering a demand node beyond the target distance is zero. However, this definition is unrealistic, because it does not differentiate between the demand nodes within the target time or distance, while it differentiates completely between the demand nodes within the target time and demand nodes which are slightly beyond the target time or distance.

The following example (Figure 1) illustrates this major problem with the traditional definition of coverage. The total area to be covered by the service is divided into smaller administrative areas or districts, for which suitable demand data is available. Assume that a station is placed at the centre of area A, and the centres of the areas A, B, C, D, E, F, G, and I are within the target time, while the other areas are beyond it. (The "centre" of the area takes into account the weighted demand, so these points are not necessarily at a geographical centre.) Assume also that the total area is a plain, and that coverage is based on the distance separating the station-area pairs. Since the traditional definition of coverage is a zero-one variable (e.g. 1 if it covered, 0 otherwise), then all demands located at these quarters within the target time are definitely covered, while demands located beyond these quarters are definitely not covered. In other words, the probability of covering A1 is the same as the probability of covering F1, and the same as the probability of covering E1 which is equal to 1, while the probability of covering L1 or K1 is zero. However, the distance separating E1 and L1 is very small compared to the distance or time separating A1 and E1. Therefore, if the probability of covering L1 is zero, then the probability of covering E1 is at least very small. In addition, if you look at L1 and F1 then you may notice that L1 is not covered while F1 is covered, even though L1 is closer than F1.

Figure 1 (about here)
Illustrating the traditional definition of coverage

A second source of error is that caused by aggregated demands. In the approaches used for the LSCP and the MCLP, demand aggregation together with the definition of coverage may give a misleading solution, Daskin et al. (1989), and Current and Schilling (1990). The binary coverage definition used in LSCP and MCLP may include or exclude demands that are on the boundary of the threshold. The errors due to demand data aggregation in the LSCP and the MCLP approaches are potentially more significant than in the *p*-median problem, not because of the problem size, but because of the definition of the coverage. However, the model that follows redefines coverage and is robust to the errors due to demand data aggregation.

In Figure 1 the demand areas located around the boundary of the area covered by station A are the main cause of the problem of aggregation. Some locations around the boundary are considered to be covered, while in fact they are not, location F1 is an example. On the other hand, demand located at K1 is located within the target time of the station A, but is theoretically not covered. This is because the demand at F1 is aggregated to its centre F, which is within the target time, and the demand at K1 is aggregated to its centre K, which is beyond the target time.

Since demands are always aggregated to finite potential areas, aggregation always exists in covering problems. However, in the model to be described the effect of aggregation is negligible, simply because demands located around the boundary of a station are giving a small weight in determining whether to locate at that station or not. The objective function of this model consists of two parts multiplied together. It maximises the demands covered multiplied by the probabilities of reaching them. The demands located very close to a potential station have high probabilities of being covered, while demands located around the boundaries of a potential station have lower probabilities of being covered. Therefore, the importance given to the demands around a potential station decreases the farther away they are. By multiplying aggregated demands by probabilities, the objective function gives more emphasis to the demands located closer to a potential station than demands located farther away. Therefore, the demands located very close will affect the choice of where to locate a station more than demands located at a greater distance or time. In other words, unlike the traditional covering problem where aggregation may affect the optimal locations, here the aggregation will not affect the optimal locations. Assume that area A is a potential station, and location F1 has about 5% of the total demands, and assume that the probability of arriving within the target time for area F is only 10 %. Using the set covering approaches, this location will add the whole 5% to the objective function. However, using the model to be described, the 5% will become only 0.5%. Therefore, though demands located within the target time from a potential station are high, their effect on the objective function depends on the distance or time separating these demands from that station.

4. Proposed Goal Programming Model

Input Variables

Input variables related to the demand in the planning region are created. For example, the city of Riyadh is divided to 92 quarters. The proportion of total demand (a_i) originating at each quarter (i = 1 to n) is used in the model. The travel times between each pair of quarters are used to determine the probability of reaching area i in the target time from station j, P_{ij}.

Decision Variables

The decision variables are the locations of the stations (j = 1 to m), the number of ambulances allocated to the stations, S, and the assignment of demand areas to their stations, Y_{ij}. Therefore, the decision for each potential station will be whether to locate there or not. Once the decision related to the locations is made, then ambulances should be allocated to each selected station.

Objective function

The objective function of the model consists of two goals. First, to maximise the expected demands covered, and second to reduce the spare capacities of located ambulances, while ensuring the minimum required performance. This is achieved by minimising the underachievement of the first goal and the over-achievement of the second goal.

Thus the objective function is

$$\text{Min } P_o d_o^- + P_1 \sum_{j=1}^{m} d_j^+ \qquad (1)$$

where

P_0: First goal.

P_1: Second goal.

d_0^- and d_j^+ are deviations.

m: the total potential locations.

Expected demands covered are calculated by the multiplication of two parts: the probabilities of covering demand areas, and the proportions of total demands which originated at these demand areas. Therefore, the first goal, P_0, can be formulated as follows:

$$\text{Max. } \sum_{i=1}^{n} \sum_{j=1}^{m} a_i P_{ij} Y_{ij}$$

where:

i: the demand areas, $i = 1$ *to* n.

j: the station areas, $j = 1$ *to* m .

n: the total number of demand areas.

m: the total number of the potential stations.

P_{ij}: the probability of reaching area i in the target time from station *j*. $P_{ij} = P_{ij}$ if it is greater than a pre-specified probability p, otherwise it is zero.

a_i: the proportion of demand originating in area *i*.

Y_{ij}: is 1 if $P_{ij} \geq \;\; p$ and station *j* is the nearest open station and can be reached within the target time; 0 otherwise. The model multiplies demands within the target time for each potential station with the probabilities of arrival on time to the demands. The highest result among the potential stations is the first chosen station, then next one is the next one chosen, and so on. This process continues until the number of open stations is the number of stations required.

This objective function can be expressed as a goal constraint in a goal programming formulation (P_0) as follows:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} a_i P_{ij} Y_{ij} + d_0^- = 1 \tag{2}$$

where

d_0^- is the under-attainment deviation. It ranges from 0 to 1, zero if $\sum_{i=1}^{n}\sum_{j=1}^{m} a_i P_{ij} Y_{ij} = 1$, which happens only if all areas are covered with a 100% probability. However, this is unlikely, especially when the available resources are limited.

Since d_0^- will be minimised, this goal constraint maximises the summation of the aggregated demands multiplied by the probabilities (i.e. the expected demands covered). In addition, since the maximum value of the expected demands covered is 1, this goal constraint is set to be equal to 1.

For the second goal, the maximum expected demands covered are fixed, and the nearest possible locations to cover these demands are known. Therefore, the first goal is now a constraint for the second objective (i.e., second goal) to determine the optimal number of ambulances that meet the performance levels. In other words, the goal is to place ambulances in each opened station in such a way that, for a pre-specified proportion of the time, any call arising within the service boundary of that station will find at least one ambulance available. This goal may be achieved by using the Erlang Loss Formula, which can be used to find the probability of having S ambulances busy in the system at the time of service request.

Erlang's Formula (M (λ) /G/S-blocking.)

$$PS = (\rho^{s} / S!) / \sum_{i=0}^{S} (\rho^{i} / i!)$$

where,

λ: arrival rate.

μ : service rate.

ρ : traffic intensity; λ/μ.

S: number of servers in the system.

PS: the probability of S servers being busy when an arrival occurs.

Using this formula the probability of having all S ambulances busy in the system at the time of service request can be found. Figure 2 shows the behaviour of the Erlang loss formula for one, two, three, and four ambulances located at a station, when the service rate is 1.67 calls per hour, as in Riyadh. It shows the curve of the probabilities of ambulances being busy for different arrival rates.

Using the Erlang loss formula, and using the expected service rate, boundary values in the arrival rates can be found. The boundary values are the arrival rates when it is necessary to increase the number of ambulances by one in order to maintain the performance level for availability of ambulances. Arrival rates are determined by the total demands assigned to a specific station. Suppose, as was the case in Riyadh, that the EMS authority wants to impose 5% as a maximum limit of the busy probability for any open centre. Table 1 shows the boundary values at which the busy probability is equal to the target, when the arrival rate, as in Riyadh, is 1.67 calls per hour.

Figure 2 (about here)

The probability of no ambulance being free for different arrival rates

assuming that the service rate is 1.67 calls per hour.

Table 1 Boundary values

Ambulances (S)	1	2	3	4
Arrival rates boundary values (r_S)	0.0875	0.636	1.497	2.541

These values may be determined by using the Newton-Raphson method of approximation and the Erlang loss formula. If α is the maximum busy probability, as specified by the decision-makers, then what is required are the values of λ which satisfy the Erlang equation, where $\rho = \lambda / \mu$, for different values of *S*.

The EMS authority for Riyadh wanted to be sure that the probability of having a busy ambulance should be at most 5% (e.g. 95% reliability). Therefore, the boundary values can be imposed in the formulation not only to ensure the reliability level but also to reduce the excess of the workload above the performance level at any opened station. The constraints set will be as follows:

$$r_S x_{jS} - \sum_{i=1}^{n} \lambda_i Y_{ij} \geq 0$$

where

r_S : the boundary value in the arrival rates from *S* to *S*+1.

x_{jS} : 1 if *S* vehicles are placed at location *j*, 0 otherwise.

λ_i : the arrival rate for node *i*.

$\sum_{i=1}^{n} \lambda_i Y_{ij}$: the overall arrival rates for the nodes served by station *j*.

By adding this constraint two things can be ensured:

1) Only those stations that are selected will have to meet the reliability constraint. If a station is not selected then the second term of the above inequality will be zero.

2) To include the actual demand covered, not all areas within the specific target time will be counted, since some of the demands may shift to another closer open station.

This approach allows the exact arrival rates and the exact numbers of ambulances to be found. The arrival rate for each area (if a station is placed in that area) will depend on the total demand areas served by that station. Other techniques used in the literature pre-calculate the minimum number of ambulances required to meet the specific performance, and, therefore, over-estimate the minimum number of ambulances required to meet the performance level(s), (Marianov and ReVelle, 1996).

The second goal can be shown in a goal programming formulation as follows:

$$\sum_{1<S<c} r_S x_{jS} - \sum_{i=1}^{n} \lambda_i Y_{ij} - d_j^+ = 0 \tag{3}$$

where

d_j^+: over-attainment or spare capacity for station j.

c: maximum number of ambulances that can be located at station j.

5. Conclusions

This paper extended models for the Maximal Covering Location Problem (MCLP) for emergency medical service in two ways. First, instead of the usual 0-1 coverage, the model has considered the more realistic situation when the probability of covering a demand within the target time varies between 0 and 1. Second, once the locations are determined, the minimum number of vehicles at each location that satisfies a specified performance level is determined. Erlang's loss formula is used to identify the arrival rates when it is necessary to add an ambulance in order to maintain the performance level for the availability of ambulances. Thus, the problem of identifying the optimal locations of a pre-specified number of emergency medical service (EMS) stations is addressed in two stages. The first goal is to locate ambulance stations so the maximum expected demand can be reached within a pre-specified target time. Then, the second goal is to ensure that any demand arising located within the service boundary of the ambulance station will find at least one ambulance available. The model developed has been applied to the Saudi Arabian Red Crescent Society (SARCS), Riyadh City, Saudi Arabia. A fuller description of that application may be found in Alsalloum and Rand (2003).

References

Alsalloum, O. I., Rand G. K., 2003. Locating ambulance stations for the EMS system of Riyadh City, Saudi Arabia. See working paper 2003/035

Ball, M. O., Lin, F. L., 1993. A reliability model applied to emergency service vehicle location, *Operations Research*, 41, 18-36.

Charnes, A., Storbeck, J., 1980. A goal programming model for the siting of multilevel EMS systems, *Socio-Economic Planning Sciences*, 14, 383-389.

Church, R. L., ReVelle, C. S., 1974. The maximal covering location problem, *Papers of the Regional Science Association*, 32, 101-118.

Current, J. R., Schilling, D. A., 1990. Analysis of errors due to demand data aggregation in the set covering and maximal covering location problem, *Geographical Analysis*, 22 (1), 116-126.

Daskin, M., Stern, E. H., 1981. A hierarchical objective set covering model for emergency medical service vehicle deployment, *Transportation Science*, 15, 137-152.

Daskin, M., 1983. A maximal expected covering location model: formulation, properties, and heuristic solution, *Transportation Science*, 17, 48-69

Daskin, M. S., Haghani, A. E., Khanal, M., Malandraki, C., 1989. Aggregation effects in maximum covering models, *Annals of Operations Research*, 18, 115-140.

Hakimi, S. L., 1964. Optimum locations of switching centers and the absolute centers and medians of a graph, *Operations Research*, 12, 450-459.

Hogan, K., ReVelle, C. S., 1986. Concepts and applications of backup coverage, *Management Science*, 32, 1434-1444.

Marianov, V, ReVelle, C. S., 1994. The queueing probabilistic location set covering problem and some extensions, *Socio-Economic Planning Sciences*, 28 (3), 167-178.

Marianov, V, ReVelle, C. S., 1995. Siting emergency services, In Drezner, Z. (ed.), Facility location: a survey of applications and methods, Springer, Heidelberg, 199-223.

Marianov, V., ReVelle, C. S., 1996. The queueing maximal availability location problem: a model for siting of emergency vehicles, *European Journal of Operational Research*, 93, 110-120.

Narasimhan, S., Pirkul, H., Schilling, D., 1992. Capacitated emergency facility siting with multiple levels of backup, *Annals of Operations Research*, 40, 323-337.

ReVelle, C. S., Hogan, K., 1988. A reliability-constrained siting model with local estimates of busy fractions, *Environment and Planning B: Planning and Design*, 15, 143-152.

ReVelle, C. S., and Hogan, K., 1989. The maximum availability location problem, *Transportation Science*, 23, 192-199.

Toregas, C., Revelle, C. S., 1972. Optimal location under time or distance constraints, *Papers of the Regional Science Association*, 28, 133-143.

Figure 1

Illustrating the traditional definition of coverage

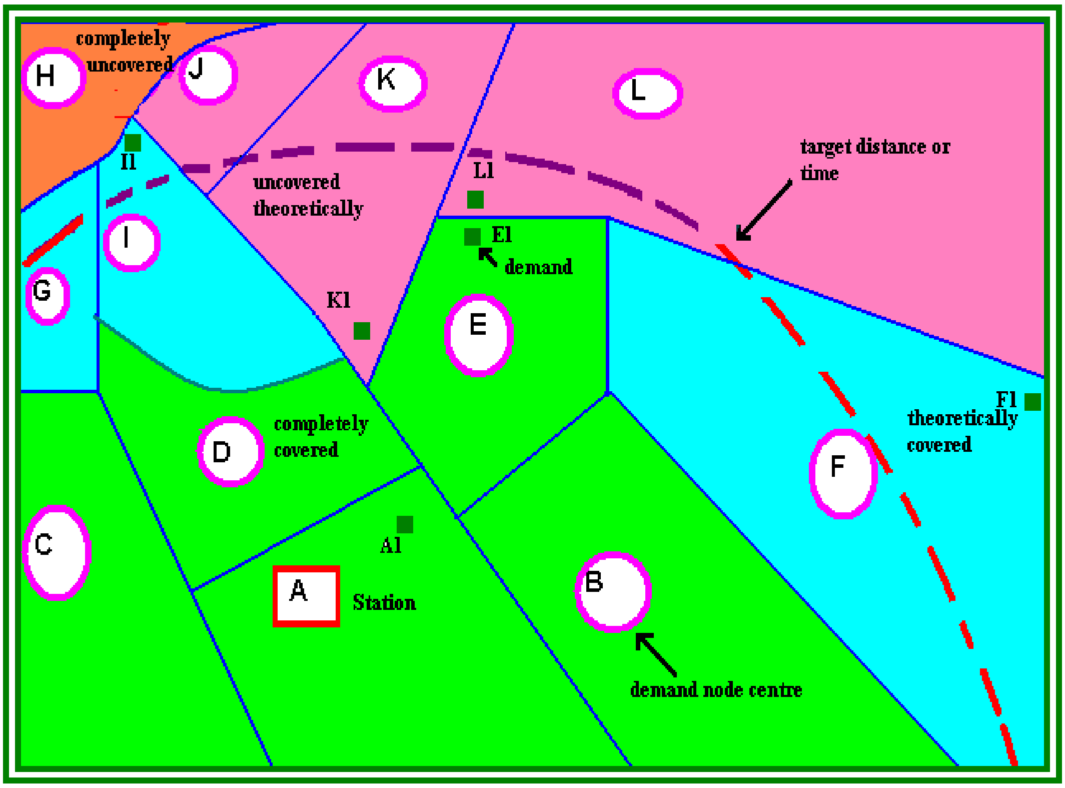
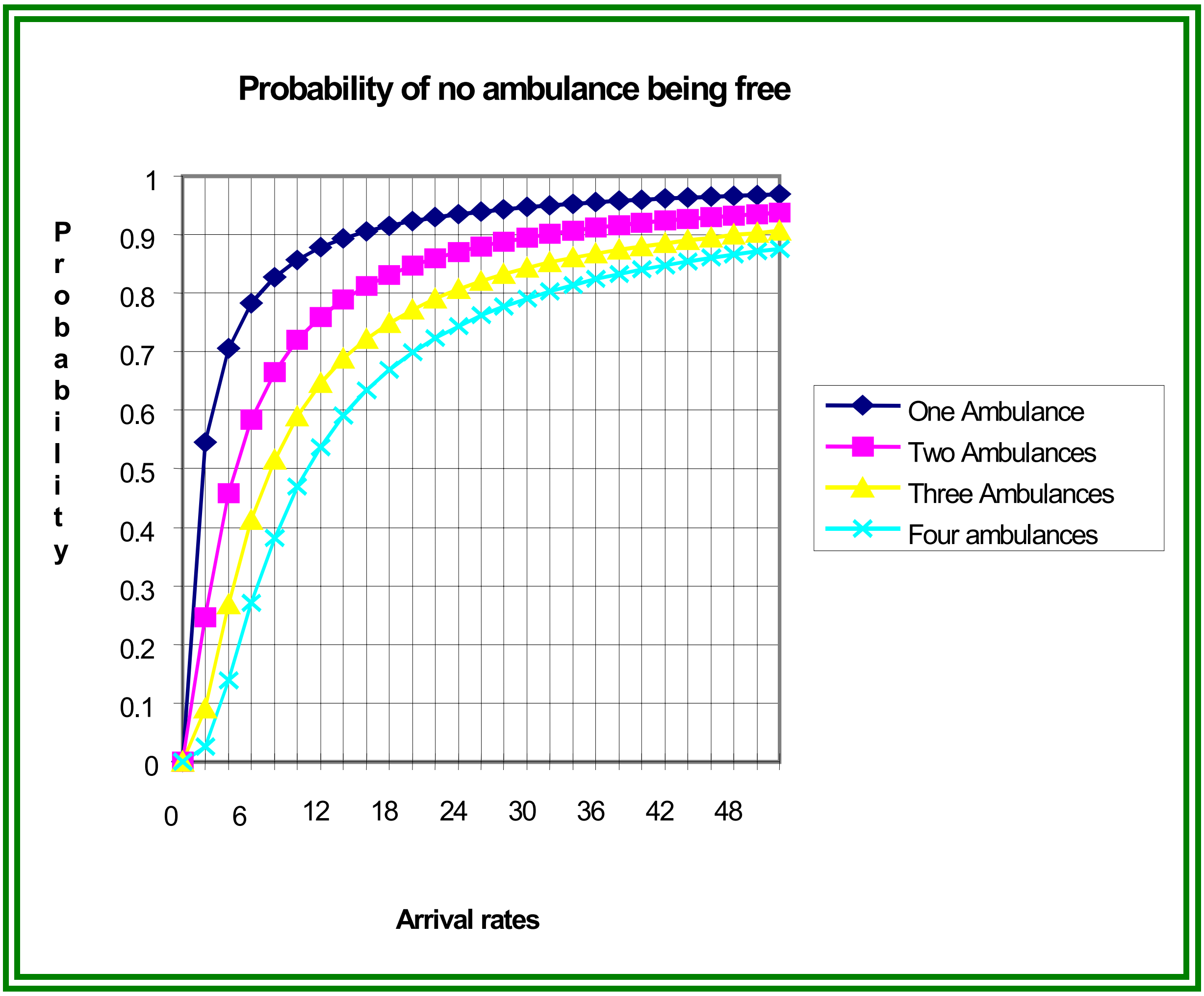


Figure 2

The probability of no ambulance being free for different arrival rates assuming that the service rate is 1.67 calls per hour.


Probability of no ambulance being free
Probability
Arrival rates
One Ambulance
Two Ambulances
Three Ambulances
Four ambulances