Extensions to emergency vehicle location models

Othman Ibraheem Alsallouma, Graham K. Randb,*

aKing Saud University, Riyadh, Saudi Arabia
bDepartment of Management Science, The Management School, Lancaster University, Lancaster LA1 4YX, UK

Available online 6 April 2005

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 that 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.
© 2005 Elsevier Ltd. All rights reserved.

Keywords: Location; Emergency medical services

1. Introduction

This paper describes a model that has been developed and applied to the emergency medical service (EMS) of Riyadh, the capital city of Saudi Arabia. The aim of the 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. Thus a model was developed to give insight to the Saudi Arabian Red Crescent Society (SARCS) as to where their EMS stations should be built and the number of ambulances to be located at each station.

The model developed here is an extension to the maximal covering location problem (MCLP), which was originally proposed by Church and ReVelle [1]. The purpose is to identify the optimal locations of a specified number of EMS stations. The first objective is to locate these stations so the maximum expected demand might be reached within a pre-specified target time. The traditional definition used in 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, this coverage definition is replaced by the probability of covering a demand within the target time. The second objective 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) that 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.

The paper is organised as follows. First, there is discussion of the literature devoted to EMS models to indicate the background of the model described in this paper, in particular set covering models and the application of such models to specific situations. Section 3 discusses the definition of coverage, introducing the concept of the probability of arriving in an area within a specified time. In Section 4, the model is developed. This can be found succinctly stated in Appendix A. The application of the model to Riyadh is described in Section 5, before some conclusions are drawn.

2. Emergency medical service models and their application

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, whereas stochastic models better address this aspect. They take into account the probabilities of servers being busy, and/or the stochastic nature of ambulances arriving at the demand points. Marianov and ReVelle [2] and Brotocorne et al. [3] present excellent surveys 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, where a 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 adapts the goal programming formulation of Charnes and Storbeck [4], and the models of Ball and Lin [5], and Marianov and ReVelle [6] to allocate the exact numbers of ambulances required in each open base.

Church and ReVelle [1] developed a model for the MCLP, which finds the location of a pre-specified number of facilities $n$ so as to maximise the demand covered by at least one facility. 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 may be busy when a call 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 model were developed. The maximum expected covering location problem (MEXCLP) developed by Daskin [7], 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 [8,9] 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 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.

Ball and Lin [5] 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. It assigns each demand to several servers, if more than one are open within the pre-set distance. This helps when deciding the minimum number of ambulances located in any opened station. The workload that Rel-P assigns to each ambulance is an upper bound on the real workload.

Marianov and ReVelle [6] 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.

The set covering location problems and their extensions have been used extensively in many real-world applications, including EMS location. Eaton et al. [10] used the MCLP model to determine optimal centres for which to recruit rural health workers and ambulances in Valle Del Cauca, Colombia, using a greedy adding and substitution heuristic. This model was also used by Eaton et al. [11] to analyse options for EMS vehicle deployment in Austin, Texas, in terms of equity, efficiency, effectiveness measures and opportunity costs associated with administrative alternatives. Later, Eaton et al. [12] applied a multiobjective formulation to determine the optimal locations of emergency medical services in Santo Domingo, Dominican Republic. In this application, the model aimed to maximise the multiple coverage, given that each node is covered at least once. The problem was solved in two steps using a multiobjective heuristic, since no integer or linear programming codes were available to the authors at that time.

Daskin’s [7] MEXCLP model was applied by Fujiwara et al. [13] to locate the EMS in Bangkok. The ‘good’ solutions obtained were further analysed by a simulation model, since the model does not give an optimal solution. Nevertheless, Fujiwara et al. [14] used MEXCLP along with the probabilistic central facility location model of Aly and White [15] to screen the large number of possible alternatives, which were also subjected to a detailed analysis by simulation.
The survey by Brotcorne et al. [3] traces the evolution of ambulance location and relocation models over the last 30 years.

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.

Fig. 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 or time separating E1 and L1 is very small compared to that 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.

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, [16,17]. 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 Fig. 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**

4.1. **Input variables**

Input variables related to the demand in the planning region are created. For example, the city of Riyadh is divided into 92 quarters. The proportion of total demand \((a_i)\) originating at each quarter \((i = 1 \ldots n)\) 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 \(j\), \(P_{ij}\).

4.2. **Decision variables**

The decision variables are the locations of the stations \((j = 1 \ldots m)\), 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.

4.3. **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 under-achievement of the first goal and the over-achievement of the second goal.

Thus the objective function is

\[
\text{Min} \quad P_0 d^-_0 + P_1 \sum_{j=1}^{m} d^+_j, \tag{1}
\]

where \(P_0\) is the first goal, \(P_1\) the second goal, \(d^-_0\) and \(d^+_j\) are deviations, and \(m\) 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 that originated at these demand areas. Therefore, the first goal, \(P_0\), can be formulated as follows:

\[
\text{Max} \quad \sum_{i=1}^{n} \sum_{j=1}^{m} a_i P_{ij} Y_{ij}, \tag{2}
\]

where \(i\) is the demand areas, \(i = 1 \ldots n\); \(j\) the station areas, \(j = 1 \ldots m\); \(n\) 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 \(j\), if it is greater than a pre-specified probability \(p\), otherwise it is zero; \(a_i\) the proportion of demand originating in area \(i\); and \(Y_{ij}\) 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 the next highest 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:

\[
\sum_{i=1}^{n} \sum_{j=1}^{m} a_i P_{ij} Y_{ij} + d_0^- = 1,
\]

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 \((M(\lambda)/G/S)\-blocking.\)

\[
PS = \frac{(\rho^S / S!)}{\sum_{i=0}^{S} (\rho^i / i!)} ,
\]

where \(\rho\) the traffic intensity, \(\lambda/\mu\); \(\lambda\) is the arrival rate; \(\mu\) the service rate; \(S\) the number of servers in the system; and \(PS\) 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. Fig. 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 that, as was the case in Riyadh, 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 service rate, as in Riyadh, is 1.67 calls per hour. These values may be determined by using the Newton–Raphson method of approximation and the Erlang loss formula. If \(z\) is the maximum busy probability, as specified by the decision-makers, then what is required are the values of \(\lambda\) 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
Fig. 2. The probability of no ambulance being free for different arrival rates assuming that the service rate is 1.67 calls per hour.

Table 1

<table>
<thead>
<tr>
<th>Boundary values</th>
<th>Ambulances ($S$)</th>
<th>Arrival rates boundary values ($r_S$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Ambulances ($S$)</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Arrival rates boundary values ($r_S$)</td>
<td>0.0875</td>
<td>0.636</td>
</tr>
</tbody>
</table>

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 $r_S$ is the boundary value in the arrival rates from $S$ to $S + 1$; $x_{jS}$ is 1 if $S$ vehicles are placed at location $j$, 0 otherwise; $\lambda_i$ the arrival rate for node $i$; and $\sum_{i=1}^{n} \lambda_i Y_{ij}$ 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.
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 arrival rates and 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), (see [6]).

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

\[
\sum_{1 < S < c} r_S x_{jS} - \sum_{i=1}^{n} \lambda_i Y_{ij} - d_j^+ = 0, \tag{6}
\]

where \(d_j^+\) is the over-attainment or spare capacity for station \(j\), and \(c\) the maximum number of ambulances that can be located at station \(j\).

5. Application to Riyadh

The EMS in Riyadh was intending to increase its service to serve a larger demand within a target time. However, the EMS authority had no idea how many stations should be added, where they should be located, and how many ambulances were needed in each station. The decision-makers of the SARCS were contacted, and the general vice president agreed that the data needed would be made available. SARCS was founded in 1983, and later became the 91st member of the International Red Cross Society. Seven ambulance stations have existed in Riyadh for a long time and each serves some designated quarters. Five of these stations are rented, while the EMS authority owns stations in quarters 17 and 21. The numbers in Fig. 3 represent the quarter identification numbers, all of which were used as potential stations. The bold lines represent the boundaries of each station, the location of which is indicated by boxes around the quarter identification numbers.

The Central Radio Communication Room (CRCR) receives the demand calls. The calls are then assigned to the station responsible for serving that area, which then dispatches one of its ambulances. If there is no available vehicle, the CRCR assigns the call to any other station, based on their judgement. Calls are served in a first come first served basis. The staff of the CRCR usually fill out a form consisting of basic data including location of the incident and times of receiving the call, receiving the call by the specific station, dispatching an ambulance, arrival at the scene, pick up time at the scene and moving to the hospital (if required to be transferred), arrival at the hospital (drop-off time) and arrival at the station. Staff of the CRCR usually do not write the time when ambulances arrive at their bases, because once an ambulance drops off a patient in hospital, this ambulance is considered ready to receive another call while on the way back to its station.

Data collection was not straightforward, as it was necessary to collect manually all the information. The EMS receives more than 13,000 calls on average a year (36 calls a day) using seven centres and about 35 ambulances. The EMS system is a self-contained system, because each station is responsible only for those demands within its boundary. Data for developing and estimating the model coefficients were collected from April 15 to July 29, 1996. The area was divided into 92 quarters, as in the CRCR records. Detailed information for more than 3800 incidents was collected. This information covers about three
months and a half. For each call the first six times in the list given earlier were collected and rounded to the nearest minute. Demand per quarter and the service time per call for each quarter were collected.

The distances or travel time between each pair of quarters needs to be estimated. A number of different ways have been suggested (see [12,13,18–24]). In this context, the probability of an ambulance located
at an existing station arriving at its service area within a specified time needs to be determined. However, the probability of an ambulance located at a potential station arriving to a service area within a specified time cannot be calculated, because of the absence of empirical data. Therefore, another way of estimating this information was needed. With the help of The Higher Commission for the Development of Arriyadh (HCDA), travel times between quarter–quarter pairs were estimated. These travel times are obtained using special computer software called EMME/2, designed for traffic and planning studies [25]. It requires a large amount of data of an existing network when used for planning. It tries to closely approximate the real-world conditions. It requires a full description of the existing network representation such as traffic surveys of all quarters, road characteristics, and accident statistics. In addition, the existing traffic characteristics of roadway or transit links such as volumes, travel times and speeds can be input for evaluation. The data of a network representation can be input by co-ordinates, or it can be digitised from a map. Up to 30 modes of travel can be input for different types of transit vehicles.

The equilibrium (capacity constrained) auto traffic assignment problem may be solved by the linear approximation algorithm. The behavioural assumption, on which the equilibrium traffic assignment problem is built, is that each driver chooses the route that he perceives to be the best. When each driver cannot improve his travel time by changing his route, then the equilibrium is achieved. In other words, “driver optimality” is achieved when no driver can improve his travel time by changing the route. The equilibrium traffic assignment corresponds to a set of flows such that all paths used between an origin-destination pair are of equal time.

The output of this software is comprehensive. It includes the travel time among all nodes within the networks. In addition, it can be used in traffic simulation models for the establishment of signal settings. Economic evaluation, traffic impact analysis, and evaluation of network performance are optional outputs. Possible outputs are practically unlimited since the software has an open architecture that allows numerous specialised analysis procedures to be built.

Using the data from the CRCR records, the travel times from each of the existing station–quarter pairs were obtained. These travel times were converted to probabilities of travelling within the target time (i.e., 10 min) for the existing station–quarter pairs, by generating cumulative curves of number of calls reached within the target times (i.e., 10 min or less). Furthermore, to make these estimations more reliable, only the station–quarter pairs that have at least 30 calls were used. These results are the dependent variables values (i.e., $y$), which will be used in the regression model to estimate the probabilities of travelling within the target time among quarter–quarter pairs for the whole city. Travel times among quarter–quarter pairs for the whole city are obtained from the EMME/2 software package. Therefore, the probability of covering a specific node was estimated in two ways. First, the probability of covering a quarter from its existing station was determined during the period of monitoring the system as described earlier. Second, the probability of covering a quarter from its potential station was estimated by combining empirical data and the results obtained by the EMME/2 software package.

Estimating probabilities on the basis of observing the probabilities between each quarter–quarter pair would have been expensive in time and resources. For example, to drive only once each quarter–quarter pair requires more than eight 1000 trips. However, to estimate the distribution of probability for every quarter–quarter pair, at least $\frac{1}{4}$ million trips are required, assuming only 30 observations per quarter–quarter pair. Therefore, a model is needed to estimate the probabilities of arriving within the specific time. Note that a model is important since no data exists regarding the probabilities of arriving within the target time between each potential station–quarter pair.
Whilst Goldberg et al. [22,23] and Goldberg and Paz [24] have used regression to estimate travel times, in this context a regression model will be used to derive an approximation for the probabilities of arriving within the specific time. The travel times are selected as independent variables for two reasons. Firstly, the travel times were the only independent variable, since it is the only reasonable data available. Secondly, the travel times incorporate all factors that affect the probabilities of arriving within the specific time. These factors are traffic conditions, speed of an ambulance, time of day, climate conditions, and type of routes. Therefore, the travel times are the only independent variables, while the probability of arriving within the specific time is the dependent variable. The travel time is the independent variable because it causes the probability of arriving to vary. The appropriate tool for finding the relationship between probability of arriving within a specific time and the travel time, which consists of other factors that affect the probability of arriving, is regression analysis.

The observed probabilities were divided into two homogeneous groups. In other words, two simple linear regressions were used to estimate the rest of the probabilities of arriving within the specific time among quarter-quarter pairs. The first group is for the existing stations located at the city centre and old quarters, which have specific similar characteristics such as narrow lanes, heavy traffic, lower speed, and so on. These stations are located at nodes 21 and 17. Therefore, the probabilities of arriving within the target time from the potential stations located within the boundary of these quarters were estimated using the result of the first regression model. The second group is for the other stations that may share or experience different characteristics such as wider streets, less traffic, higher speed, and so on. These stations are located outside the city centre. Similarly, the probabilities of arriving from the potential stations located within the boundary of these stations were estimated using the result of the second regression model.

While a linear regression has been used in this context, it is important to consider alternative relationships. These data were tested against linear, logarithmic, inverse, quadratic, cubic and other regression models to see which describes the curve best. Once the best model is found, the second step is to use this model in estimating the relationship. If a non-linear model is appropriate to the actual data, then the probabilities of travel within the specified time can be estimated from the model that has been determined to be the best.

Fig. 4 shows a comparison between the probability of arriving within the target time by the first linear regression model and the conventional set covering models. It shows how the probability of covering a demand decreases steadily as the arriving time increases. It starts with the intercept (0.93), and decreases by $-0.058$ for each unit time increase. On the other hand, the set covering models assume that any demand within the target time (i.e., 10 min) is 100% covered, while others beyond the target time are covered with 0% probability.

The results that are obtained reflect the behaviour of the model. As far as the first goal is concerned, the nodes that can access areas of high demands and with higher probabilities of being accessed will be targeted. The selection of a node to be a base is determined by the value that this selection can add to the objective function. This model tries to increase not only the total demands covered, but also the probability of being covered. Therefore, the selected node must increase both these two factors, or if that is impossible, then increase one of them.

Writing the formulation for such a large problem was a major task and the nature of the problem, which is a binary (zero–one) integer-programming problem, made it difficult to solve optimally. These obstacles were tackled in two ways. First, to solve the problem optimally, the goals were split, some constraints transformed, and the size of the problem reduced to optimise the second goal. Second, to write the
The formulation itself, a coefficient matrix and Pascal codes were used. The following three approaches were applied to solve the problem to optimality.

The first and second goals were split into two problems: first to find the optimal locations and second to find the fewest ambulances required to satisfy the performance levels. Before the problem constraints were transformed, the problem was solved by linear programming to see which variables turned out to be integers and which variables turn out to be fractions. Surprisingly, all variables turn out to be either one or zero except the constraints of the following type:

\[ X_j - \sum_{i=1}^{n} Y_{ij} > 0. \]

All \(X_j\) values turned out to be fractions, which means that the problem of solving optimally arises from this type of constraints. Therefore, the above constraints were changed to the following form:

\[ X_j - Y_{1j} > 0 \cdots X_j - Y_{nj} > 0. \]

As a result, the problem becomes larger due to the expansion in the total number of constraints. Before the transformation the number of the constraints was less than 300, but after the above transformation there were about 1500 constraints. However, this transformation allowed the problem to be solved in seconds. After the transformation, the problem was solved for one location up to 17 locations (covering the whole city), and all the runs finished optimally within 2 s.

The second goal is to find the minimum number of ambulances required in each opened location to satisfy the performance levels. In this case, the approach used for the first goal is not applicable, since the aggregated demand rates assigned to each opened station should be included in one constraint. Therefore, another way of solving the second goal optimally is required. The only straightforward way is to reduce the size of the problem, since the optimal locations are already found in the first goal. In other words, there are 92 potential locations in the first goal, but in the second goal the optimal locations found will be included while the others will be removed from the problem. Therefore, in the second goal the size of the problem is reduced after excluding the variables and constraints related to the non-optimal locations. Therefore,
Table 2
The optimal locations and their objective values

<table>
<thead>
<tr>
<th>No. of locations</th>
<th>Locations</th>
<th>Maximum expected demands covered</th>
<th>% of population covered</th>
<th>Average probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22</td>
<td>0.301</td>
<td>0.448</td>
<td>0.672</td>
</tr>
<tr>
<td>2</td>
<td>22,54</td>
<td>0.385</td>
<td>0.566</td>
<td>0.680</td>
</tr>
<tr>
<td>3</td>
<td>22,54,69</td>
<td>0.464</td>
<td>0.657</td>
<td>0.706</td>
</tr>
<tr>
<td>4</td>
<td>22,54,69,10</td>
<td>0.538</td>
<td>0.742</td>
<td>0.725</td>
</tr>
<tr>
<td>5</td>
<td>54,69,8,14,57</td>
<td>0.592</td>
<td>0.810</td>
<td>0.730</td>
</tr>
<tr>
<td>6</td>
<td>54,69,8,14,57,42</td>
<td>0.629</td>
<td>0.873</td>
<td>0.721</td>
</tr>
<tr>
<td>7</td>
<td>14,22,47,54,57,42,69</td>
<td>0.659</td>
<td>0.853</td>
<td>0.770</td>
</tr>
<tr>
<td>8</td>
<td>14,22,31,47,54,57,42,69</td>
<td>0.681</td>
<td>0.861</td>
<td>0.790</td>
</tr>
<tr>
<td>9</td>
<td>14,22,31,47,54,57,42,69</td>
<td>0.702</td>
<td>0.867</td>
<td>0.810</td>
</tr>
<tr>
<td>10</td>
<td>14,22,36,31,47,54,57,42,69</td>
<td>0.722</td>
<td>0.887</td>
<td>0.814</td>
</tr>
<tr>
<td>11</td>
<td>14,22,36,31,47,54,57,42,69</td>
<td>0.740</td>
<td>0.918</td>
<td>0.806</td>
</tr>
<tr>
<td>12</td>
<td>14,22,36,31,47,54,57,42,69</td>
<td>0.756</td>
<td>0.918</td>
<td>0.823</td>
</tr>
<tr>
<td>13</td>
<td>14,22,36,31,47,54,57,42,69</td>
<td>0.772</td>
<td>0.934</td>
<td>0.826</td>
</tr>
<tr>
<td>14</td>
<td>14,22,36,31,47,54,57,42,69</td>
<td>0.787</td>
<td>0.948</td>
<td>0.828</td>
</tr>
<tr>
<td>15</td>
<td>14,22,36,31,47,54,57,53,43,84,48,69,63,92</td>
<td>0.799</td>
<td>0.965</td>
<td>0.828</td>
</tr>
<tr>
<td>16</td>
<td>14,22,36,31,47,54,57,53,43,84,48,69,63,92</td>
<td>0.809</td>
<td>0.978</td>
<td>0.827</td>
</tr>
<tr>
<td>17</td>
<td>14,22,36,31,47,54,57,53,43,84,48,69,63,86,92</td>
<td>0.818</td>
<td>0.990</td>
<td>0.827</td>
</tr>
</tbody>
</table>

the optimal solutions for the second goal are obtained within seconds, even though the constraints, which caused a problem in the first goal, are included, because of the smaller size of the problem.

The optimal solutions obtained for 1–17 locations, the number required to cover the entire city, are given in Table 2. The demand density map given in Fig. 5 will aid interpretation of these results.

For a single station node 22 is selected. There are only two major superhighways that divide the city into four sections: King Fahad Road and Makkah road. Node 22 is located at the intersection of these two major highways. From this node the superhighways allow quick access to most of the city. Therefore, it has been chosen to cover most of the demands within the target time. Furthermore, it is optimal because it is the only place that covers two highly populated clusters. One is in the south towards the city centre (i.e., nodes 1–21), and the other cluster is in the north of the city centre (i.e., nodes 36, 37, 39, 40 and 57).

From the results of the model, the existing seven stations only cover 74% of the population within 10 min. If SARCS wish to implement these results they have at least three options: cover the whole city, relocate the existing stations, or increase the number of stations to give a specific service level. To cover the whole city, at least 17 stations are needed to satisfy the minimum requirements that any node can be served from a station within at most 10 min and within the maximum possible probability of arriving. Twenty-four ambulances are needed to ensure that any call that arises anywhere at the city will find at least an idle ambulance within 10 min driving to serve it. Relocating the existing stations will maximise the demands within the target time and will maximise the probability of serving them on time. By relocating the existing seven stations to these quarters, the demands covered will be about 85% of the population, which is an improvement of 11%. Two ambulances are required at each of these locations. Several quarters are not covered within the target time, but can be assigned to the nearest opened stations. By assigning the uncovered quarters to the nearest opened stations, then the arrival rates will be increased. Therefore,
the number of ambulances allocated to each opened base may need to increase too, to keep the same level of performance (i.e., the probability of an ambulance being busy when a call occurs is less than 5%). However, by knowing the assignment of the uncovered quarters, the new allocation of ambulances can be calculated. The results obtained allow the number of stations to give a specific service level to be determined. Alternatively, the seven existing stations, though not optimal, can be regarded as fixed...
and optimal solutions, given this constraint, can be found for 8–18 stations, which in this scenario are required to cover the entire city. The results are presented in Table 3.

The results were discussed with SARCS. They requested an investigation of a specific location, whether it is optimal or whether there is a better one. SARCS were interested in locating at a specific building at Almalaz (quarter 26). However, the exact location of the building or street (i.e., within Almalaz quarter) is not dealt with explicitly by the model. The authority of the EMS will decide which building is suitable and available to be a station. If the EMS is given the opportunity to locate anywhere in a quarter, then the existing street network and intersections need to be examined to find the point that minimises shortest paths to all demands covered. However, in reality it is easier to locate somewhere in a quarter but it is more difficult to locate at an exact point. There are many reasons that make it more practical for a model to locate at a quarter instead of location at a specific point. For instance, there may not be a building at the exact point, or if there is a building it may be not suitable because of its size or facilities. In addition, even though you find a suitable building at the exact point, it may not be available to be rented or to be bought. Therefore, the model does not allow finding the exact point, so cannot be used to suggest whether to locate at a specific building or not. However, the model can indicate whether locating at Almalaz is optimal or not.

By adding an additional station at Almalaz (e.g., as a quarter) to the existing seven stations, the maximum expected coverage will increase from 0.559 to 0.60. However, by locating at Ghobirah (numbered 10), as seen in Table 3, the maximum expected coverage will increase from 0.559 to 0.623. Ghobirah is the eighth optimal location, which covers the Southeast of the city centre. The area around it is a populated area and is relatively far from the existing stations.

There is a logic for the model not locating at Almalaz (numbered 26), simply because there is a large overlap between Almalaz and the existing station Almoraba (station numbered 21). Manfuha (numbered 12) or Aryan (numbered 44) are two locations which are optimal at this level of coverage, and are robust locations, since they are included in the set of nine locations and will be optimal until the city is covered.

### Table 3
Summary of the optimal locations (by fixing the existing stations)

<table>
<thead>
<tr>
<th>No. of locations</th>
<th>Locations</th>
<th>Maximum expected demands covered</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>17,21,54,81, 48, 69, 63</td>
<td>0.559</td>
</tr>
<tr>
<td>8</td>
<td>17,21,54,81,48,69,63,10</td>
<td>0.623</td>
</tr>
<tr>
<td>9</td>
<td>17,21,54,81,48,69,63,12,44</td>
<td>0.678</td>
</tr>
<tr>
<td>10</td>
<td>17,21,54,81,48,69,63,12,44,47</td>
<td>0.706</td>
</tr>
<tr>
<td>11</td>
<td>17,21,54,81,48,69,63,12,47,42,31</td>
<td>0.728</td>
</tr>
<tr>
<td>12</td>
<td>17,21,54,81,48,69,63,12,31,47,42,53</td>
<td>0.747</td>
</tr>
<tr>
<td>13</td>
<td>17,21,54,81,48,69,63,12,26,31,47,42,53</td>
<td>0.765</td>
</tr>
<tr>
<td>14</td>
<td>17,21,54,81,48,69,63,12,26,31,47,42,53,92</td>
<td>0.780</td>
</tr>
<tr>
<td>15</td>
<td>17,21,54,81,48,69,63,12,26,31,47,42,53,92</td>
<td>0.793</td>
</tr>
<tr>
<td>16</td>
<td>17,21,54,81,48,69,63,12,26,31,47,53,43,84,92</td>
<td>0.805</td>
</tr>
<tr>
<td>17</td>
<td>17,21,54,81,48,69,63,12,26,31,47,53,43,84,92,51</td>
<td>0.815</td>
</tr>
<tr>
<td>18</td>
<td>17,21,54,81,48,69,63,12,26,31,47,53,43,84,92,51,86</td>
<td>0.825</td>
</tr>
</tbody>
</table>
completely. In other words, by locating at either Manfuha or Aryan you can ensure that this location will still be an optimal station as long as the distributions of demands persist. Choosing between Manfuha and Aryan depends on your preference. By locating at Manfuha (numbered 12), the maximum expected coverage will increase from 0.559 to 0.619, but if locating at Aryan (numbered 44), the maximum expected coverage will increase from 0.559 to 0.615.

One of the authors visited SARCS some three years after the work was completed. The locations of the stations are still as they were because of budgetary considerations, and because they could not re-locate the very old locations to better sites because of long-term rental agreements for five of the sites. However, they adopted another way of maximising the expected coverage of demands within a specified period of time. An ambulance has been located in each optimal location on standby. These ambulances are equipped to deal with many cases at the scene of the incident, without the need to rush the patient to hospital.

6. 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: first goal to locate ambulance stations so the maximum expected demand can be reached within a pre-specified target time, and then to ensure that any demand arising located within the service boundary of the ambulance station will find at least one ambulance available. The results were applied to the EMS in Riyadh, Saudi Arabia.

Appendix A. The goal programming model

The whole model can be expressed as

Objective function

\[
\text{(1)} \quad \text{Min } P_0 d_0^- + P_1 \sum_{j=1}^{m} d_j^+, \tag{A.1}
\]

where \( P_0 \) is the first given priority, first goal; \( P_1 \) the second given priority, second goal; \( d_0^- \) and \( d_j^+ \) are deviations; and \( m \) the total number of potential locations.

This objective function minimises the total deviations of two things. First, to minimise \( d_0^- \) means to maximise the expected demands covered by the closest facility. Second, to minimise \( d_j^+ \) means to reduce the excess above the pre-specified performance level.
Subject to

Goal constraints:

\[ \sum_{i=1}^{n} \sum_{j=1}^{m} a_i P_{ij} Y_{ij} + d_0^- = 1 \]  \hspace{1cm} (A.2)

and

\[ \sum_{1 < S < c} r_S x_{jS} - \sum_{i=1}^{n} \lambda_i Y_{ij} - d_j^+ = 0. \]  \hspace{1cm} (A.3)

System constraints:

\[ \sum_{j=1}^{m} Y_{ij} \leq 1. \]  \hspace{1cm} (A.4)

This constraint assigns each demand area to the nearest open station only, which has a higher probability of reaching this demand within the pre-specific target time limit. Otherwise, it is not covered.

\[ \sum_{S=1}^{c} x_{jS} \leq 1, \]  \hspace{1cm} (A.5)

where \( S \) is the number of ambulances located at station \( j \), which will only be selected if \( j \) is selected to be a station; and \( c \) the maximum number of ambulances that can be located at station \( j \).

This constraint either places a specific number (\( S \)) of ambulances in station \( j \) or does not place any ambulance there (i.e., station \( j \) is not optimal).

\[ \sum_{j=1}^{m} \sum_{S=1}^{c} x_{jS} = p, \]  \hspace{1cm} (A.6)

where \( m \) is the total potential stations, and \( p \) the maximum number of stations that can be opened.

References


