# Multi-server Queuing Maximum Availability Location Problem with Stochastic Travel Times

Noraida A. Ghani

***Abstract*— This paper extends the Queuing Maximum Availability Location Problem (Q-MALP) model to locate ambulances offering two kinds of services, i.e., Advance Life Support (ALS) and Basic Life Support (BLS). The development of the model includes the randomness of server availability and of travel times. This model is applied to a 33-node census tract representation of Austin, Texas. The implications of the new model for the ambulance services system design are discussed as well as the limitations of the modeling approach.**

***Index Terms*— Ambulance deployment, set covering location problems, stochastic travel time, multi-server**

## I. Introduction

A critical component of the emergency health care system is a responsive and well-managed ambulance service. An almost universal measure of ambulance location system performance is the response time, generally defined as the time between dispatch of the emergency medical personnel until arrival of the personnel on-scene. Dispatching time depends on dispatching policies and the delay is independent of the ambulance location. The only component that is affected by changing the location of the ambulance is the travel time. Thus, for ambulance location, using travel time as a surrogate for response time is a common practice and is the most meaningful measure [1]-[3].

Ambulance services systems often face many types of demand for service and provide multiple levels of emergency care. Demands on ambulance services can be broadly grouped into two categories: critical and non-critical. The former comprises calls of a potentially "life-threatening" nature while the latter describes calls which are considered emergent but "non-life threatening". Ambulance services can also be grouped into two broad categories: advanced life support (ALS) and basic life support (BLS). ALS service is provided by paramedic units equipped to effectively handle critical demands such as airway management and cardiac resuscitation. BLS service is provided by emergency medical technician (EMT) units equipped to respond to less urgent and non-critical problems. Although EMT units are not equipped with ALS capabilities, a role exists for these units in responding to critical calls. In place of immediate paramedic assistance, the EMT units can perform first aid and basic life-support procedures (oxygen, control of external bleeding and other interventions) in critical situations until ALS-equipped vehicles arrive.

The objective of this study is to develop a technique for locating ambulances in an urban environment. First, we will extend the Queuing Maximum Availability Location Problem (Q-MALP) model to locate two types of ambulances, i.e., the BLS and the ALS. Henceforth, this model will be known as the Multi-server Queuing Maximum Availability Location Problem (MQ-MALP). Next, we will extend the formulation when the stochastic nature of the urban environment is taken into account explicitly, namely, the travel time in response to service. A measure of uncertainty of the response time, i.e., a probability measure, will be incorporated into the existing optimization model.

## II. The Notion of Coverage in the MQ-MALP Model

The ambulance location system is represented as a network of nodes and arcs. The nodes of the network constitute demand points or demand areas as well as potential locations for the ambulance units. The arcs that connect nodes to one another are feasible routes, usually taken to be the shortest travel times between the nodes. Demand at a particular node is measured by the number of calls/time unit and if unavailable, by it's proxy, the population size. The maximum (standard) travel time and the number of available ambulance units to be located will be provided by the decision maker. Of particular importance in these models is the set $N_{si}$ defined as follows:

$N_{si} = \{ j \mid t_{ij} < S, j \in J \}$ where $J$ = set of eligible facility sites, $t_{ij}$ = shortest time from potential facility site $j$ to demand node $i$ and $S$ = time standard for coverage of critical calls. In other words, $N_{si}$ is the set of facility sites located within the time standard S of demand node $i$. If a call for service originating at this node is answered by available ambulances stationed inside this neighborhood, it will be answered within the time standard.

In the Q-MALP model, a demand/call for service is considered "covered" if there is an ambulance available within the time standard with the stated α-reliability. With this notion of "coverage", the Q-MALP is now extended to locate 2 types of ambulances, i.e., the BLS and ALS units. The coverage of two different types of calls, then, suggests two travel time standards, $S$ for covering critical calls and $T$ for covering non-critical calls ($S < T$). First, coverage of

Manuscript received March 2, 2012; revised April 12, 2012.

N. A. Ghani is with the School of Distance Education, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia (phone: 604-653-3636; fax: 604-657-6000; e-mail: noraida@usm.my).

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

critical calls by paramedics and EMT (ALS or BLS) units is considered. We wish to deploy these ambulances (ALS or BLS) units in such a way as to cover this critical demand within $S$ time units with an ambulance available with reliability α. Further, since ALS units are more appropriately equipped for this task, we seek to maximize primarily their coverage of these critical calls; BLS units are serving as only a "back-up" function. Next, we also wish to deploy BLS units in such a way as to cover the non-critical demand within T time units with the BLS unit available with reliability α. Thus, the BLS units in the above system are assigned two tasks, that of providing "back-up" coverage of critical demand and that of providing coverage of non-critical demand.

The use of "coverage" as defined above requires an estimate of the probability that an ambulance is busy or sometimes referred to as busy fraction. The estimate presented here is analogous to the Q-MALP in which we not only use region-specific busy fractions (i.e., server busy fraction is adjusted to the call level around demand node) but we also allow dependence between busy fractions at a local, neighborhood level. These assumptions provide an improvement over the total independence assumptions or the situation in which the probability of the server being busy is the same across the whole system (as in Maximum Expected Covering Location Problem (MEXCLP)).

### *A. Deriving server needs with reliability $\alpha$ for ALS or BLS units*

A call of critical nature is considered "covered" if there is an ALS or BLS server within S time standard with reliability α. To achieve such reliability a minimum number of ALS or BLS units ($b_{ai}$) must be derived, located within $S$ time standard of node $i$, such that the probability that any of these servers are busy is less than 1-α.

The demand of critical nature in each neighborhood $S$ of demand node $i$ is modeled as an $M/G/s/s$ system [4], i.e., the number of critical calls for service in neighborhood $S$ of demand node $i$ is assumed to be distributed as Poisson arrivals with intensity $\lambda_{ai}$ and the service time served by the ALS or BLS unit is assumed to be general with a mean rate of $\mu_{ai}$. When all the ALS servers in the neighborhood are busy, new calls are presumed lost (servers from outside the neighborhood take these calls). Implicit in this model is the assumption [see, e.g., [5]] that the flows of servers into neighborhood S of demand node $i$ and out of neighborhood $S$ of demand node $i$ approximately cancel each other and thus justify the treatment of each neighborhood as an isolated, independent unit whose demands and servers interact solely with each other.

Let $s_i$ be the total number of ALS or BLS servers in the neighborhood $S$ of $i$ and define the state $k$ of the ambulance system with critical calls as $k$ ALS or BLS servers being busy within the neighborhood of demand node $i$. Using the standard queuing theory steady-state equations, the probability $p_k$ of the system being in state $k$ can be computed as [4]:

$P$(getting into state $k$) – $P$(getting out of state $k$)

$= \left(\lambda_{ai}\ p_{k-1} + (k+1)\,\mu_{ai}\ p_{k+1}\right) - \left(\lambda_{ai}\ p_k + k\,\mu_{ai}\ p_k\right) = 0$,

for states 1, 2,..., $s_i$

where

$p_k$ = the probability that $k$ servers within the neighborhood $S$ of demand node $i$ are busy,

$\lambda_{ai}$ = arrival rate of critical call in neighborhood $S$ of demand node $i$ (calls per day),

$1/\mu_{ai}$ = mean service time of an ALS or BLS unit (hours per call) responding to a call in the neighborhood of demand node i , and

$\mu_{ai}\, p_1 - \lambda_{ai}\, p_0 = 0$, for state 0.

For convenience, the subscript $i$ is omitted from $p_k$. At steady-state, solution of these equations yields the probability of all $s_i$ servers being busy, $p_{si}$ [4]:

$$p_{si} = \frac{(1/s_i!)\rho_{ai}^{s_i}}{1+\rho_{ai}+(1/2!)\rho_{ai}^2+\ldots+(1/s_i!)\rho_{ai}^{s_i}}, \text{ where } \rho_{ai} = \mu_{ai}/\lambda_{ai} \quad (1)$$

The parameters $\lambda_{ai}$ can be estimated by $\sum_{k\in M_{si}} f_{ak}$ and the parameters $\mu_{ai}$ can be estimated by $24/\overline{t_a}$ (calls per day) where

$f_{ai}$= demand of critical natureat demand node (number of calls per unit time) and

$\overline{t_a}$ = average duration of a single critical call serviced by an ALS or BLS unit (in hours) averaged over the entire network.

$M_{si} = \{\ k \mid t_{ik} < S,\ k \in I\ \}$ where $I$ = set of demand notes,

$t_{ik}$ = shortest time from demand node $i$ to demand node $k$ and $S$ = time standard for coverage of critical calls. In other words, $M_{si}$ is the set of demand nodes located within $S$ of $i$.

The recursive formula for $p_{si}$ is given by [4]:

$$p_{s_i} = \left(\frac{1}{p_{s_i-1} + s_i\mu_{ai}/\lambda_{ai}}\right) p_{s_i-1}. \quad (2)$$

Since the term in the parenthesis is less than 1, $p_{si}$ then is a decreasing function of the parameter $s_i$. Expressions (1) and (2) are identical to the expression in the original Q-MALP Model [see [5]].

For each neighborhood $S$ of demand node $i$ and each value of $s_i$, the probability of at least one ALS or BLS server being available is 1- $p_{si}$ . If the value 1- $p_{si} \geq \alpha$ or equivalently, $p_{si} \leq 1 - \alpha$, then demand node $i$ is assumed to be covered with reliability α. As $p_{si}$ is a decreasing function of $s_i$, there exist a nonnegative integer $b_{ai}$, such that for $s_i \geq b_{ai}, 1 - p_{si} > \alpha$ [5]. The integer $b_{ai}$ represents the minimum number of ALS or BLS units which must be located within $S$ time standard of demand node $i$, for that node to be considered as covered with reliability α. That is, $b_{ai}$ is the smallest integer satisfying

$$\frac{(1/b_{ai}!)\rho_i^{b_{ai}}}{1+\rho_{ai}+(1/2!)\rho_{ai}^2+\ldots+(1/b_{ai}!)\rho_{ai}^{b_{ai}}} \leq 1-\alpha. \quad (3)$$

The value of $b_{ai}$ can be calculated by determining $p_1$, $p_2$, ..., $p_{si}$, $p_{si+1}$, ..., etc. in sequence, and choosing $b_{ai}$ as the smallest value of $s_i$ that satisfies (3). Thus, to maximize critical calls with α-reliable service, we maximize critical calls with $b_{ai}$ or more servers. Given a value for α and knowing the values of $\lambda_{ai}$ and $\mu_{ai}$, integer $b_{ai}$ can be pre-

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

computed or determined exogenously to the optimization problem.

### B. Deriving server needs with reliability α for BLS units

Further, let $b_{bi}$ represent the minimum number of BLS units which must be located within $T$ unit of demand node $i$ for node $i$ to be covered with reliability α. Similarly, given a value for α, $b_{bi}$ can be pre-computed using (3) by changing the values of $\rho_{ai}$ to $\rho_{bi}$ where

$$\rho_{bi} = \lambda_{bi} / \mu_{bi},$$

and

$\lambda_{bi}$ = arrival rate of non-critical call in neighborhood $T$ of demand node $i$ (calls per day),
$1/\mu_{bi}$ = mean service time of a BLS unit (hour per call).

That is, $b_{bi}$ is the smallest integer satisfying

$$\frac{(1/b_{bi}!)\rho_i^{b_{bi}}}{1+\rho_{bi}+(1/2!)\rho_{bi}^2+...+(1/b_{bi}!)\rho_{bi}^{b_{bi}}} \le 1-\alpha. \quad (4)$$

The parameters $\lambda_{bi}$ can be estimated by $\sum_{k \in M_{Ti}} f_{bk}$ and the parameter $\mu_{bi}$ can be estimated by $24/\bar{t}_b$ (calls per day) where

$f_{bi}$ = demand of non-critical nature at demand node i (number of calls per unit time)and
$\bar{t}_b$ = average duration of a single non-critical call serviced by BLS unit (in hours) averaged over the entire network.
$M_{Ti} = \{k \mid t_{ik} < T, k \in I\}$ where $I$ = set of demand notes, $t_{ik}$ = shortest time from demand node $i$ to demand node $k$ and $T$ = time standard for coverage of non-critical calls. In other words, $M_{Ti}$ is the set of demand nodes located within $T$ of $i$.

An algorithm that calculates minimum server needs with reliability α (smallest $b_{ai}$ as in (3) and smallest $b_{bi}$ as in (4)) was developed in Visual C++. This algorithm was applied to the 33-node problem representing Austin, Texas with a few modifications and is presented in the next section.

### C. Example analysis of server needs for 33-node case study

The 33-node problem from Daskin [6] represents Austin, Texas, at the census tract level. Interzonal travel times are given by travel matrix with intrazonal times taken to be one minute. The weights associated with each zones are the number of calls for ambulance services recorded in the census tract during the five-months period for which the data were available. However, for the purpose of analysis, the population concentration at each node was multiplied by a constant factor such that the resulting average calls per day over the entire network is 0.4 calls per day. These are used as estimates of the number of calls per node per day. Second, we used the finding from Eaton *et al*. [7] - only 20-25 percent of calls for ambulance service in Austin require the advanced skills of paramedics - in order to get estimates for the breakdown of these calls into critical and non-critical calls. Thus, the population concentration at each node was multiplied by a constant factor such that the resulting average calls for ALS services per node per day over the entire network is 0.1 calls per day. These are used as estimates of the number of critical calls per node per day. The difference between these two estimates are used as estimates of the number of non-critical calls per node per day. Third, an average duration of a single service of 3/4 of an hour was used [5]. This figure was estimated based on the average of three cases: the ambulance goes to the site of the call, stays there for some time, and then goes back to the facility site; the ambulance reaches the emergency site, takes the patient to a hospital and returns to its assigned facility site; and the possibility of a false alarm, or the event that the emergency is over when the ambulance reaches the alarm site. Lastly, the response time was set at 8 and 10 minutes for an ALS unit and a BLS unit, respectively.

Table I shows the minimum number of servers needed at varying levels of reliability, for different set of scenarios. For example, the second entry (second column, fourth row) in Table I specifies that 20 nodes require1 BLS unit and 13 nodes require 2 BLS units to achieve 85% server reliability with a response time standard of 10 minutes. The third column provides similar analysis focusing on ALS units but with a smaller response time standard of 8 minutes to reflect the nature of a critical call.

Table I. Minimum number of ALS and BLS units required under different reliabilities for the33-node problem

| Reliability | BLS Response Time 10 #locations (#servers) | ALS Response Time 8 #locations (#servers) |
|---|---|---|
| 0.80 | 33(1) | 33(1) |
| | - | - |
| 0.85 | 20(1) | 33(1) |
| | 13(2) | - |
| 0.90 | 12(1) | 33(1) |
| | 21(2) | - |
| 0.95 | 2(1) | 3(1) |
| | 31(2) | 30(2) |
| 0.99 | 16(2) | 5(1) |
| | 17(3) | 28(2) |

It can be seen that the number of locations with a higher value of minimum number of servers needed at each node in order for the node to be considered as covered with reliability α, increases with an increase in server reliability. This is intuitive as one would expect that increasing the number of ambulances at each location would increase the likelihood of an ambulance continuing to be available within the standard even after one of the ambulances had responded to a call.

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

## III. MQ-MALP with Travel Time Uncertainty Model Formulation

One way to reduce costs while maintaining quality service is to design a system in which personnel with less training would be dispatched in lower-cost BLS vehicles to non-life threatening events. Highly trained paramedics in ALS vehicles would respond to life-threatening (critical) calls. In this section we are concerned with locating a limited number of ambulances (ALS or BLS) units. The objective is to provide a maximal cover of critical calls by ALS or BLS units and a maximal cover of non-critical calls by BLS units. Thus, the BLS units in this system are assigned two tasks, that of providing "back-up" coverage of critical demand (calls) and that of providing coverage of non-critical demand (calls).

### A. Objective function

Since $b_{ai}$ ALS or BLS ($b_{bi}$ BLS) servers (pre-calculated in previous section) are required by each demand area for α-reliable coverage of critical (non-critical) calls, a lack of a full complement of servers leads to a lack of α-reliable coverage of critical (non-critical) calls. Hence, the objective to maximize the population (or calls) coverage of critical (non-critical) nature with α-reliable service, the population with the full complement, $b_{ai}$ ($b_{bi}$), of servers needs to be maximized. We now define

$$y_{aik} = \begin{cases} 1, \text{if at least } k \text{ ALS or BLS units are within } S \text{ of demand node } i \\ 0, \text{otherwise, } k = 1, 2, ..., p^a + p^b, \end{cases} \text{ and}$$

$$y_{bik} = \begin{cases} 1, \text{if at least } k \text{ BLS units are within } T \text{ of demand node } i \\ 0, \text{otherwise, } k = 1, 2, ..., p^b, \end{cases}$$

where

$p_a$ = number of available ALS units to locate, and
$p_b$ = number of available BLS units to locate.

With these definitions, the objective of maximizing coverage of critical calls by ALS or BLS units can be formulated as

$$max \sum_{i \in I} f_{ai} y_{aib_{ai}},$$

while the objective of maximizing coverage of non-critical calls by BLS can be formulated as

$$max \sum_{i \in I} f_{bi} y_{bib_{bi}}.$$

If these two objectives are used in a multiple programming problem, these objectives then becomes

$$maximize\ w_a \sum_{i \in I} f_{ai} y_{aib_{ai}} + w_b \sum_{i \in I} f_{bi} y_{bib_{bi}}, \qquad (5)$$

where strictly positive weights $w_a$ , $w_b$ represent the tradeoffs amongst the two objectives. Weights need to be strictly positive to insure noninferiority, since some alternate optima maybe inferior [8].

### B. Constraints

In this formulation, the fact that the $k^{th}$ ALS unit should not be located at a facility node without the $(k-1)^{st}$ ALS unit, is enforced through the ordering constraints. Thus, analogous to Q-MALP, yaik cannot be one unless $y_{ai(k-1)}$ is also one and the same relationship holds for $y_{bik}$. These contraints are

$$y_{aik} \le y_{ai(k-1)} \quad \forall_i,\ k = 2, 3, \ldots, b_{ai}, \qquad (6)$$

and

$$y_{bik} \le y_{bi(k-1)} \quad \forall_i,\ k = 2, 3, \ldots, b_{bi}, \qquad (7)$$

Further, to count coverers for each demand node, we define

$$x_{aj} = \begin{cases} m, \text{if at least } m \text{ ALS servers are located at site } j, \\ 0, \text{otherwise, } m = 1, 2, ..., p^a. \end{cases}$$

and

$$x_{bj} = \begin{cases} m, \text{if at least } m \text{ BLS servers are located at site } j, \\ 0, \text{otherwise, } m = 1, 2, ..., p^b. \end{cases}$$

Expressions $\sum_{j \in N_{si}} x_{aj}$ and $\sum_{j \in N_{si}} x_{bj}$ represent the total number of ALS and BLS servers that are stationed within $S$ of node $i$, respectively. In order for node $i$ to be covered $b_{ai}$ times, there would have to be at least $b_{ai}$ ALS or BLS servers that are stationed within $S$ of node $i$. Thus constraint

$$\sum_{k=1}^{b_{ai}} y_{aik} \le \sum_{j \in N_{si}} x_{aj} + \sum_{j \in N_{si}} x_{bj} \qquad \forall_i \in I, \qquad (8)$$

defines coverage for critical demand by ALS and the "back-up" coverage of critical demand by BLS units, that is, node $i$ is covered $b_{ai}$ times only if at least $b_{ai}$ ALS or BLS servers are stationed within $S$ of node $i$. Similarly, expression $\sum_{j \in N_{Ti}} x_{bj}$ represents the total number of BLS servers that are stationed within $T$ of node $i$. In order for node $i$ to be covered $b_{bi}$ times, there would have to be at least $b_{bi}$ BLS servers that are stationed within $T$ of node $i$. Thus, constraint

$$\sum_{k=1}^{b_{bi}} y_{bik} \le \sum_{j \in N_{Ti}} x_{bj} \qquad \forall_i \in I, \qquad (9)$$

defines coverage for non-critical demand by BLS units, i.e., node $i$ is covered $b_{bi}$ times only if at least $b_{bi}$ BLS servers are stationed within $T$ of node $i$.

Often the number of available ALS and BLS units are limited due to, for example, budget constraint. Expressions $\sum_{j \in J} x_{aj}$ and $\sum_{j \in J} x_{bj}$ represent the total number of ALS and BLS units that are located in the system. Hence, to limit the number of available ALS units to locate to $p_a$ and the number of available BLS units to locate to $p_b$, constraints

$$\sum_{j \in J} x_{aj} = p_a, \qquad (10)$$

$$\sum_{j \in J} x_{bj} = p_b, \qquad (11)$$

should be included in the model formulation. Further, the number of servers that can be located at a specific location is usually constrained by its capacity. The capacity $C_j$ (in servers) of each location $j$ can be reflected in the model as constraints

$$x_{aj} + x_{bj} \le C_j \quad \forall j \in J, \qquad (12)$$

and

$$x_{aj}, x_{bj} = integer \le C_j \quad \forall j \in J, \qquad (13)$$

Finally, constraints

$$y_{aik}, y_{bik} = 0, 1 \qquad \forall i, k \qquad (14)$$

forces coverage variables to be binary (i.e. 0, 1).

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

### C. Travel Time Uncertainty

Travel time is an important component of response time and is the most directly affected by deployment changes. It is important to note that travel times are random; that is they cannot be predicted exactly in advance. Even if the ambulance traveled from a particular facility site to the same street corner over and over again under essentially constant conditions-same vehicle, driver, weather, time of day, etc.- there would still be variations in travel time from run to run. If the conditions changed between runs there would be even greater variations. Thus, the random nature of travel time must be taken into account when doing analysis using travel time as a performance measure.

In the formulations of Q-MALP the response time/travel times along the arcs of the network are assumed to be deterministic. In other words, the probability distribution of the response time is degenerate. Daskin [9] formulated his location, dispatching and routing model by treating travel times as normally distributed with known mean and variance. While assuming normally distributed travel time makes his analysis considerably more tractable, it risks losing some realism because normal distribution admits the possibility of negative travel times. Marianov and ReVelle [5] on the other hand, proposed a different approach to the treatment of travel times but still using the same distribution. However, no analysis was done to see the effect of random travel time on server location. Abdul Ghani [10] and Abdul Ghani and Mohd Ruslim [11] on the other hand, uses a Monte Carlo simulation of travel times and of demand, respectively, as inputs into the optimization model with a heuristic method developed to site the ambulances.

Analogous to the approach by Marianov and ReVelle [5], we will use other distribution, namely the Weibull distribution. Suppose now the response times $T_{ij}$ are non-degenerate random quantities with probability distribution, $F_{Tij}$ . By treating the response times as random quantities, an improvement can be introduced in the way $N_{si}$ is computed. This can be done by choosing a neighborhood of each node in such a way that, if a call for service originating at this node is answered by an available server located within the neighborhood, it will be answered within time standards with probability $\gamma$ . To do this, $N_{si}$ is redefined as

$$N_{Si} = \left\{ j \middle| P(T_{ij} \le S) \ge \gamma \right\} = \left\{ j \middle| F_{T_{ij}} ) \ge \gamma \right\}. \tag{15}$$

When $F_{T_{ij}}^{-1}$ exists, then (15) can be rewritten as

$$N_{Si} = \left\{ j \middle| S \ge F_{T_{ij}}^{-1}(\gamma) \right\} = \left\{ j \middle| F_{T_{ij}}^{-1}(\gamma) \le S \right\}. \tag{16}$$

If travel time is distributed as Weibull with shape parameter $\alpha_{ij} > 0$, and scale parameter $\beta_{ij} > 0$, then (16) becomes

$$N_{Si} = \left\{ j \middle| \beta_{ij}(-\ln(1-\gamma))^{1/\alpha_{ij}} \le S \right\}.$$

A full formulation of MQ-MALP is given in Tables II, III and IV.

Table II. Input constant

| | | |
|---|---|---|
| $I$ | = | set of demand nodes (indexed by $i$). |
| $J$ | = | set of eligible facility sites (indexed by $j$). |
| $t_{ij}$ | = | shortest time from potential facility site $j$ to demand node $i$. |
| $S$ | = | time standard for coverage of critical calls. |
| $T$ | = | time standard for coverage of non-critical calls. |
| $f_{ai}$ | = | demand of critical nature at node $i$ (number of calls per day). |
| $f_{bi}$ | = | demand of non-critical nature at node $i$ (number of calls per day). |
| $\alpha$ | = | reliability of a server. |
| $b_{ai}$ | = | the minimum number of ALS or BLS units which must be located within $S$ unit of node $i$ for node $i$ to be covered with reliability $\alpha$, pre-computed using (3). |
| $b_{bi}$ | = | the minimum number of BLS units which must be located within $T$ unit of node $i$ for node $i$ to be covered with reliability $\alpha$, pre-computed using (4). |
| $p_a$ | = | number of available ALS units to locate. |
| $p_b$ | = | number of available BLS units to locate. |
| $c_j$ | = | capacity of site $j$. |

$w_a$ , $w_b$ , $\ge 0$ are the weights associated with each objective.

Table III. Decision variables

$$y_{aik} = \begin{cases} 1, \text{ if at least } k \text{ ALS or BLS units are within } S \text{ of demand node } i, \\ 0, \text{ otherwise,} \end{cases}$$

$$y_{bik} = \begin{cases} 1, \text{ if at least } k \text{ BLS units are within } T \text{ of demand node } i, \\ 0, \text{ otherwise.} \end{cases}$$

$$x_{aj} = \begin{cases} m, \text{ if at least } m \text{ ALS servers are located at site } j, \\ 0, \text{ otherwise,} \end{cases}$$

$$x_{bj} = \begin{cases} m, \text{if at least } m \text{ BLS servers are located at site } j, \\ 0, \text{ otherwise.} \end{cases}$$

Table IV. MQ-MALP Formulation with Travel Time Uncertainty

$$Max\ Z = w_a \sum_{i \in I} f_{ai} y_{ai} b_{ai} + w_b \sum_{i \in I} f_{bi} y_{bi} b_{bi}$$

$s.t.$

$$\sum_{k=1}^{b_{ai}} y_{aik} \le \sum_{j \in N_{Si}} x_{aj} + \sum_{j \in N_{Si}} x_{bj} \qquad \forall i \in I,$$

$$\sum_{k=1}^{b_{bi}} y_{bik} \le \sum_{j \in N_{Ti}} x_{bj} \qquad \forall i \in I,$$

$$y_{aik} \le y_{ai(k-1)} \qquad \forall i, k = 2,3,...,b_{ai},$$

$$y_{bik} \le y_{bi(k-1)} \qquad \forall i, k = 2,3,...,b_{bi},$$

$$\sum_{j \in J} x_{aj} = p^a,$$

$$\sum_{j \in J} x_{bj} = p^b,$$

$$x_{aj} + x_{bj} \le C_j \qquad \forall j \in J,$$

$$x_{aj}, x_{bj} = \text{integer} \le C_j \qquad \forall j \in J,$$

$$y_{aik}, y_{bik}, = 0, 1 \qquad \forall i, k.$$

$$N_{si} = \left\{ j \middle| t_{ij} \le S, j \in J \right\}, N_{Ti} = \left\{ j \middle| t_{ij} \le T, j \in J \right\}.$$

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

## IV. Analysis of MQ-MALP Model with Travel Time Uncertainty

This section considers the application of the MQ-MALP with stochastic travel times to the 33-node problem representing Austin, Texas with the same modifications as outlined in Section C. The travel time distributions are Weibull distributions with scale and shape parameters, $\alpha_{ij}$ and $\beta_{ij}$, chosen such that the means equalled to their counterparts for the deterministic response times. A constant variance of 4 were used for all travel times. For the 33-node problems, reliability for server availability and reliability for the response times were set at 0.95. The numbers of ALS and BLS servers to locate were set at 2 and 6, respectively. While for the 55-node problem, reliability for server availability and reliability for the response times were set at 0.95 and 0.90, respectively, and the number of ALS and BLS servers to locate were set at 3 and 8, respectively. Results of these analyses were obtained using Xpress-MP version 13.1. Using the method of Cohon et al. [12], the solution procedure results are summarized in Table V. For this table, otherwise indicated by an asterik (*), only one ALS or BLS server is to be located at the indicated nodes.

The use of the method by Cohon et al. [12] resulted in several alternative solutions that do well under both objectives and that a range of potential performance levels make themselves known. For example, in moving from solution A to solution C, percent of critical calls covered increases by 19.8 percent but percent of non-critical calls covered decreases by a relatively smaller amount of 8.9 percent. "Covered" for critical calls here is defined as having an available server, either an ALS or a BLS, responding to a critical call within 8 minutes with probability of at least 0.95 and the probability of the server being available to respond is at least 0.95.

Table V. Weighting Method Results of MQ-MALP Uncertainty Model
Applied to the 33-Node Problem

| Solution | | Percent Call Covered | | | Node Location | |
|---|---|---|---|---|---|---|
| | W | Critical | Non-critical | All | ALS | BLS |
| A | 0.38 | 50.4 | 68.7 | 64.1 | 8, 13 | 6*, 15*, 24* |
| B | 1.35 | 57.9 | 66.7 | 64.5 | 8, 20 | 6, 7, 14, 15, 25* |
| C | 2.75 | 70.2 | 59.8 | 62.4 | 23, 27 | 2, 6, 8, 14, 15, 20 |
| D | 17.40 | 71.7 | 52.1 | 57.0 | 23, 29 | 2,7,10, 14, 20, 27 |
| E | 39.70 | 74.7 | 33.6 | 43.9 | 23, 27 | 2,8,14, 15, 20, 29 |

* 2 BLS servers located

Fig. 1 below shows the spatial distribution of servers corresponding to solutions in Table V. One can see that the location of the two ALS servers in the 33-node problem are more dispersed in solution C and E, while in solution D the two ALS servers are more concentrated in the middle and two BLS servers located on the same location (node 25). Which solution the decision maker picks would depend on the kind of tradeoffs he/she is willing to make. However, the spatial distribution of the more dispersed location of the servers are intuitively more appealing. In addition to improving the public sense of safety by more proximate locations, the system itself may perform better by potentially decreasing the mean response time in the dispersed formation. Furthermore, the more dispersed siting pattern could enhance the ability of the system to continue to respond should key access routes from the consolidated locations be blocked.

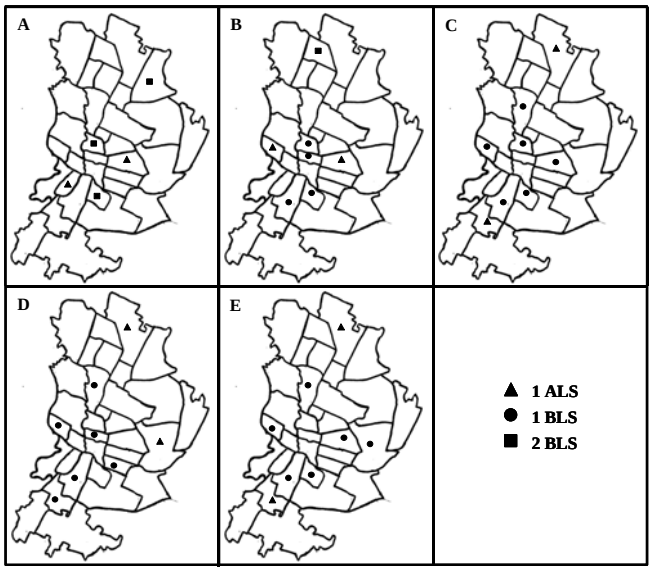


Fig. 1. Spatial Distribution of Servers Corresponding to Solutions of MQ-MALP Uncertainty Model for the 33-Node Problem in Table V.

## V. Conclusions and Recommendations

The MQ-MALP formulated here is an attempt to account for the uncertainties in the travel times in the context of locational planning of the ambulance services systems. These systems are characterized by multiple levels of emergency health services, i.e., a two-tiered system with "basic life support" and "advanced life support". The use of the multiobjective format in the formulation allows the decision maker to identify good location alternatives and the tradeoffs between them. The model integrates several of the important determinants of the emergency system performance into a single optimization model in order to strategically locate and allocate the ambulances. While the formulation progresses the state of the art of the emergency services modeling, the problem solution and the methodology presented herein are not intended to be definitive due to their limitations.

First, the queueing method, introduced by Marianov and Re Velle [5] and adapted herein to accomodate multiple servers (i.e. a BLS unit backing up an ALS unit), of obtaining the minimum number of servers to achieve the stated reliability does not take into account the possibility that these minimum number could be placed at a site that satisfies two or more demand nodes. Hence, it underestimates the number of servers needed to achieve the stated reliability. Incorporation of this possibility into the model could be an area of future research. Second, this model as in the case of Q-MALP, does have a tightly defined objective. That is, coverage for a demand node is achieved only when sufficient servers are located to achieve a response within the time standard with reliability $\alpha$. A

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

demand node is not counted as covered if there is one server less than the sufficient number required to achieve the specified level of service. Since this model maximizes calls with sufficient servers, some demand nodes may be left with very few servers within the standard in order to shift more demand nodes to the category of "sufficient". This situation can be rectified by the introduction of two other constraints, i.e., basic coverage and goal coverage as was used by Groom [13] and also proposed by Marianov and Re Velle [5]. The basic coverage would be the minimum level at which all demand nodes must be served. While maintaining basic coverage, goal coverage would seek to maximize the number of calls to the desired level. The addition of these constraints, however, would come at the expense of a decreased objective value. Third, the model do not explicitly consider other stochastic nature that are often important in designing emergency service such as demand for services. Hence, incorporating the stochastic demand into the formulation would be another important extension. Finally, it may be the case that some areas are strategically more important than others, such as areas where schools are located. In order to incorporate nodes that are strategically important, the optimization model can be adapted by adding weights to those nodes.

## REFERENCES

[1] Burt, M. John Jr. and J. S. Dyer, "Estimation of travel times in multiple mode systems," *Operational Research Quarterly*, Vol. 22, No. 2, pp. 155-163, 1974.

[2] Chelst, Kenneth and J. P. Jarvis, "Estimating the probability distribution of travel times for urban emergency service systems," *Operations Research*, Vol. 27, No. 1, pp. 199-203, 1979.

[3] K. L. Rider, "A parametric model for the allocation of fire companies in New York City," *Management Science*, Vol. 23, No. 2, pp. 146-158, 1976.

[4] D. Gross and C.M. Harris, *Fundamentals of Queuing Theory*. 3rd ed. New York: John Wiley & Sons, 1998.

[5] V. Marianov and C. ReVelle, "The queuing maximal availability location problem: A model for the siting of emergency vehicles," *European Journal of Operational Research*, Vol. 93, pp. 110-120, 1996.

[6] M. S. Daskin, "Application of an expected covering model to emergency medical service system design," *Decision Sciences*, Vol. 13, pp. 416-439, 1982.

[7] D. Eaton, M. Hector, V. Sanchez, R. Lantigua and J. Morgan, "Determining ambulance deployment in San Domingo, Dominican Republic," *Journal of the Operational Research Society*, pp. 113, 1985.

[8] J. Cohon, *Multiobjective Programming and Planning*. New York: Academic Press, 1978.

[9] M. S. Daskin, "Location, dispatching and routing models for emergency services with stochastic travel times," in *Spatial Analysis and Location-Allocation Models*, A. Gosh and G. Ruston, Ed. New York: Van Nostrand Reinhold, 1987.

[10] N. Abdul Ghani, "Siting of ambulances in a network with interdependent stochastic travel times," presented at the 13th National Mathematical Science Symposium, Alor Setar, Kedah, Malaysia, 31 May – 2 Jun 2005.

[11] N. Abdul Ghani and N. Mohd Ruslim "An application of the p-Median problem under uncertainty in demand in emergency medical services," in *Proceedings of The 2nd IMT-GT 2006 Regional Conference on Mathematics, Statistics and Applications*, Penang, Malaysia, 2006, pp. 244-250.

[12] J. Cohon, R. Church and D. Sheer, "Generating multiobjective trade-offs: An algorithm for bicriterion problems," *Water Resources Res.*, Vol. 15, pp. 1001-10010, 1979.

[13] K. Groom, "Planning emergency ambulance services," *Operational Research Quarterly*, Vol. 28, pp. 641-651, 1977.

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)