

# Scenario-based approach for the ambulance location problem with stochastic call arrivals under a dispatching policy

Inkyung Sung<sup>1</sup> · Taesik Lee<sup>1</sup>

© Springer Science+Business Media New York 2016

**Abstract** This paper proposes a scenario-based ambulance location model which explicitly computes the availability of ambulances with stochastic call arrivals under a dispatching policy. The model utilizes two-stage stochastic programming to represent the temporal variations in call arrivals as a set of call arrival sequences. Constraints are embedded in the model to ensure that available ambulances are assigned to incoming calls following a dispatching policy. A logic-based Benders decomposition algorithm is presented to solve the model. The advantage of using our algorithm is demonstrated by comparing its performance with those of other location models.

**Keywords** Ambulance location problem · Stochastic programming · Stochastic call arrivals · Dispatching policy · Logic-based Benders decomposition

## 1 Introduction

The ambulance location problem arises because of the need to determine the best locations to base ambulances so as to minimize the time required to respond to emergency medical service (EMS) requests. The location of available ambulances is a major factor determining the response times to call arrivals. Extensive research has been performed in this regard since as early as the 1970s (Church and Velle 1974; Toregas et al. 1971).

Ambulance location problems are often formulated as covering problems. A demand site is considered covered if it can be reached from an ambulance station

---

✉ Taesik Lee  
taesik.lee@kaist.edu

Inkyung Sung  
inkyung@kaist.ac.kr

<sup>1</sup> Department of Industrial and Systems Engineering, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea

within a specified time standard. The problem involves finding the optimal number of and locations for ambulances such that the sum of the covered demand sites is maximized. The covering problem consists of two major parts: the location set covering problem (LSCP) and the maximal covering location problem (MCLP). LSCP minimizes the number of facilities that are required to cover all demands, whereas MCLP maximizes the number of covered demands with a given number of facilities.

Another class of location problems is the location-allocation problem, which determines the locations of ambulances subject to a constraint that all demands are assigned to ambulances. One of the actively studied problems in this class is the  $P$ -median problem, which locates ambulances and allocates all the service demands to the ambulances such that the total (or average) travel time experienced by  $p$  ambulances is minimized (Hakimi 1964). Another stream of the location-allocation problem is the  $P$ -center problem, which locates  $p$  ambulances to minimize the maximum coverage distance (or time) while covering all demands.

The classic ambulance location models for both the covering and location-allocation problems were based on the deterministic approach. These models assume that ambulances are always available to respond to emergency calls. With the assumption, a demand is considered covered if at least one ambulance is located within a certain time standard, or can be assigned to any ambulances.

Recently, the ambulance location models were extended to incorporate the stochastic characteristics of an EMS system. In particular, much research has been conducted to take into account the availability of ambulances by using the concept of a busy fraction, i.e., the probability of being unavailable to respond to a call. From relatively simple calculations to queueing theory, various models have been developed to properly estimate the busy fraction of ambulances (Marianov and ReVelle 1996; Pereira et al. 2015; ReVelle and Hogan 1989). Based on predetermined busy fractions, the models either calculate the expected value of coverage or use a chance constraint to represent the probability of at least one ambulance being available to serve a particular request.

However, it is difficult to assume or approximate appropriate values for the busy fraction. The busy fraction is an output of the location model and cannot be known a priori. Moreover, to analytically model the busy fraction, most of the work in respect of busy fraction models assumes stationary call arrivals and uses the average call arrival rate in the models, whereas actual EMS call data shows that the volume of call arrivals varies significantly during the course of a day, on weekdays versus weekends, and between seasons (Matteson et al. 2011).

In this study, as an alternative, we propose an ambulance location model which explicitly computes the availability of an ambulance by determining when an ambulance is dispatched in response to an incoming call and when it is available to serve next call (when it returns to its home station). Specifically, given a call arrival data, which contains information for when and where a call is arrived, the proposed model determines ambulance location and dispatching decisions so as to maximize the number of covered calls. It allows us to make a strategic level of decision (i.e., ambulance locations) considering decisions made during the process of EMS operations (i.e., ambulance dispatching to calls).

We construct the model by applying a stochastic programming approach. Stochastic programming is a solution framework for solving an optimization problem with two types of decisions, *here-and-now* and *recourse*. A here-and-now decision is a proactive and planning decision that should be made before observing specific outcomes (e.g., production cost or future demands). A recourse decision is made in reaction against the observations on the outcome and depends upon the here-and-now decision made earlier. In our problem, an ambulance location and a dispatching decision correspond to the here-and-now and recourse decisions, respectively.

In the stochastic programming approach, the uncertainties are represented as a set of scenarios (i.e., possible future). With the decision structure, stochastic programming derives a solution that performs well across all scenarios. In our problem, a scenario is defined by a sequence of call arrivals. Then, we derive ambulance location and dispatching solutions that maximize the number of covered calls across all possible call arrival sequences.

It should be noted that, in the model, the ambulance dispatching decisions to demands are determined following a specific dispatching policy. A dispatching policy determines which of the ambulances available at a given moment is sent to serve an incoming call. The choice made for the current call determines the available ambulances and their coverage for the next call that arrives. This implies that there is an interaction effect between the location decision and dispatching policy; an optimal location solution under one dispatching policy may not be optimal for another policy. Therefore, ambulance locations should be determined by considering an ambulance dispatching policy. Our model is designed to explicitly incorporate the effect of a dispatching policy on the solution of the ambulance location problem.

The remainder of this paper is structured as follows. In Sect. 2, we present related work on scenario-based ambulance location models. Section 3 describes the proposed model for the ambulance location problem. Solution methods for the model are discussed in Sect. 4. Then using EMS call data for the city of Daejeon in Korea, we demonstrate the performance of the proposed model in Sect. 5, followed by a detailed discussion in Sect. 6. Finally, we conclude our paper in Sect. 7.

## 2 Related literature

Since the early 1970s, various ambulance location models have been developed ranging from a relatively simple, deterministic model—set covering/ $P$ -median/ $P$ -center—to probabilistic models—availability/reliability models—, and to dynamic models—the relocation/redeployment problem. The large volume of literature in this area of research has compelled us to review only the most closely related papers describing the application of a scenario-based model to ambulance location problems. We refer readers to Brotcorne et al. (2003), Daskin and Dean (2005), Farahani et al. (2012), Jia et al. (2007), Li et al. (2011), Owen and Daskin (1998), ReVelle and Eiselt (2005) for a comprehensive review of ambulance location problems.

We first review the literature that incorporates scenarios involving deterministic ambulance location models. To the best of our knowledge, Schilling (1982) was the first to attempt to apply a scenario-based approach to the ambulance location problem. The study extended the models for the LSCP and MCLP by incorporating scenarios and a common facility concept, where the latter is defined as a facility determined to be opened for all possible scenarios. The proposed models were designed to maximize the number of common facilities and covered demands. In Jia et al. (2007), the authors applied the scenario-based approach to the MCLP, the  $P$ -median, and  $P$ -center problems. These authors used an objective function to minimize the expected *regret* associated with a scenario, which is a measure for the difference in the objective values of the optimal solution for the scenario and the compromising solution. In the scenario-based approach, the best compromising solution that performs well across all scenarios may not be the optimal solution for a particular scenario. Carson and Batta (1990) relied on the use of scenarios to represent changing daily demand conditions and suggested the relocation of ambulances for each scenario to ensure that the total response time is minimized. The  $P$ -median problem was applied to each scenario to find the best ambulance locations.<sup>1</sup> Nickel and Reuter-Oppermann (2016) presented a two-stage program based on a simple set covering formulation, which minimizes total cost associated with base locations and ambulances. In Nickel and Reuter-Oppermann (2016), uncertainties in the number of ambulance requests at demand sites were represented as a set of scenarios.

A scenario-based approach has also been applied to probabilistic ambulance location models. For example, Beraldi and Bruni (2009) proposed a stochastic programming model for ambulance locations, which incorporates probabilistic constraints to determine the reliability of the coverage. The constraints ensure that the probability of each demand point being covered by an ambulance is greater than a certain threshold level. The model uses the constraints to determine the locations of ambulances so as to minimize the expected total travel cost (time) over all scenarios and fixed cost for opening ambulance sites. In addition, Huang and Fan (2011) applied the scenario-based approach to the maximum availability location problem (MALP), in which a demand was considered covered if a multiple number of ambulances are located within a specified time standard. The number for each demand point is determined by a simple reliability constraint (ReVelle and Hogan 1989). In Huang and Fan (2011), the numbers for a demand point were varied among scenarios. The model uses the setting to determine ambulance locations to maximize the expected coverage for all scenarios.

In this paper, we aim to provide a model which determines an ambulance location solution by using stochastic programming. While the approach we take in our work follows the prior scenario-based ambulance location models, the contributions of this study can be noted as follows. First most of the work on scenario-based ambulance location models is based on the classical ambulance location models (e.g., MALP). It is known that these models have problems with

---

<sup>1</sup> In a strict sense, the model of Carson and Batta (1990) is not a scenario planning model because the location decisions for each scenario are not linked.

respect to properly estimating the availability of ambulances. Instead of using exogenously assumed busy fraction, we make ambulance assignment decisions within our model, thereby allowing the model to compute the exact availability of ambulances. Next, the existing ambulance location models ignore or simplify a dispatching policy which has an impact on the availability of ambulances. In our model, a dispatching policy is explicitly incorporated to assign an available ambulance to an incoming call. This allows us to derive an ambulance location solution under a consideration of a target EMS system in a more precise fashion. Following sections present the model and solution techniques in detail.

### 3 Problem statement

Consider the problem of locating  $p$  ambulances to maximize covered demands in a region. In this problem we assume that uncertainties arise in the arrival of temporal and geographical demands. Let us represent the uncertainties as a set of scenarios, in which case our objective would be to determine the ambulance locations that perform well across all scenarios.

Let  $\xi$  denote a random vector for call arrivals with a support  $\Xi$  and known distribution  $P$ . We assume that  $\xi$  has a finite support, and there are  $N$  realizations  $\xi^r$ ,  $r \in \{1, \dots, N\}$ . A realization of the random vector is considered as a scenario. The stochastic program for the ambulance location problem can be written as follows:

$$\max \left\{ \sum_{r=1}^N p_r \cdot f(\mathbf{x}, \xi^r) : \mathbf{x} \in X \subseteq \mathbb{Z}^+ \right\}, \tag{1}$$

where  $\mathbf{x}$  is a decision vector for ambulance locations,  $p_r$  is the probability of scenario  $r$ , and  $f(\mathbf{x}, \xi^r)$  is the sum of demands covered by solution  $\mathbf{x}$  under scenario  $\xi^r$ .

In this paper, a scenario is defined as a sequence of call arrivals, which contains the temporal and geographical information of the calls. We let  $D^r$  denote the set of call arrivals in scenario  $\xi^r$ , indexed by  $d \in \{1, \dots, |D^r|\}$ . In the set, the calls are sorted according to their arrival time,  $a_d$  such that a call that arrives earlier has a smaller index. Set  $V$  refers to a set of all stations we can locate an ambulance at. We let  $x_d^j$  denote the integer variable indicating the number of available ambulances at station  $j \in V$  when call  $d$  arrives. The ambulance dispatching decision is denoted by a 0–1 integer variable  $y_d^j$ , which equals 1, if an ambulance located at  $j$  station is dispatched to call  $d$ ; otherwise,  $y_d^j$  equals 0. We use  $W_d^j$  to specify whether call  $d$  can be covered by station  $j$ .  $W_d^j$  equals 1 if travel time between call  $d$  and station  $j$  is less than a specified time standard for coverage; otherwise,  $W_d^j$  equals 0. Thus  $W_d^j \cdot y_d^j$  represents the coverage of call  $d$ .  $W_d^j \cdot y_d^j$  equals 1 if an ambulance located at station  $j$  is dispatched to the call and the ambulance can arrive at the call  $d$ 's location within the time standard; otherwise  $W_d^j \cdot y_d^j$  equals 0.

As noted earlier, the dispatching decision  $y_d^j$  is determined by a chosen dispatching policy. The most evident dispatching policy is the nearest-available

policy. Under the nearest-available policy, a dispatcher sends an ambulance from the station nearest to the incoming call. Other dispatching policies include a jurisdiction-based dispatching or likelihood dispatching policy. To incorporate a dispatching policy in the optimization model, we introduce a priority set  $L_d^j$  for each demand. Given a dispatching policy,  $L_d^j$  is a set of ambulance stations to serve demand  $d$  with higher priority (responsibility) than station  $j$ . For example, with the nearest-available dispatching policy,  $L_d^j$  contains all ambulance stations that are located closer to demand  $d$  than ambulance station  $j$ . In addition to representing a dispatching policy in the model, we also need to keep track of availability of ambulances at each station. To do that, we introduce a demand set  $A_d^j = \{e \in D^r \mid a_{d-1} < a_e + R_e^j \leq a_d\}$ .  $a_e$  is the arrival time of call  $e$  and  $R_e^j$  is the time to serve call  $e$  by an ambulance dispatched from station  $j$ . That is,  $A_d^j$  is a list of ambulances that are expected to return to—hence become available at—station  $j$  before a dispatching decision for an imminent call  $d$  is made.

By using this notation, we propose a two-stage stochastic programming formulation for (1) as follows:

$$\max \sum_{r=1}^N p_r \cdot f(\mathbf{x}_0, \zeta^r) \tag{2}$$

$$\text{s.t. } \sum_{j \in V} x_0^j \leq p, \tag{3}$$

$$x_0^j \in \mathbb{Z}^+ \quad \forall j \in V, \tag{4}$$

$$\text{where } f(\mathbf{x}_0, \zeta^r) = \max \sum_{d \in D^r} \sum_{j \in V} W_d^j \cdot y_d^j \tag{5}$$

$$\text{s.t. } \sum_{j \in V} y_d^j \leq 1 \quad \forall d \in D^r, \tag{6}$$

$$y_d^j \leq x_d^j \quad \forall d \in D^r, j \in V, \tag{7}$$

$$x_d^j = x_{d-1}^j - y_{d-1}^j + \sum_{e \in A_d^j} y_e^j \quad \forall d \in D^r, j \in V, \tag{8}$$

$$p \cdot y_d^j \geq x_d^j - p \cdot \sum_{i \in L_d^j} x_d^i \quad \forall d \in D^r, j \in V, \tag{9}$$

$$x_d^j \in \mathbb{Z}^+ \quad \forall d \in D^r, j \in V, \tag{10}$$

$$y_d^j \in \{0, 1\} \quad \forall d \in D^r, j \in V. \tag{11}$$

Before describing the formulation we introduce two assumptions. First, ambulances on their way back to their home station are not available to respond to a next service

request until they have returned to the station. This seems to be a standard operational practice for most ambulance systems. It should also be pointed out that chance of such event is most likely very slim unless an ambulance system is tremendously overloaded. Second, a call that arrives at a moment when all ambulances are busy is lost. If this happens, instead of waiting for a next ambulance to become available, patients will find other means of transportation such as self-transport or public transportation. Given a typical level of utilization for ambulance systems, chance of all ambulances being busy is also very slim under normal circumstances.

The first stage of the problem (2–4) determines the ambulance locations to maximize the expected number of covered demands for all scenarios. Note that  $x_0^j$  defines the ambulance location at the beginning of a planning horizon; hence, it is the solution for our location problem. The solution is determined subject to constraint (3), which limits the total number of ambulances to be located at  $p$ .

The second stage of the problem (5–11) involves the recourse decisions,  $x_d^j$  and  $y_d^j$ . Once the location decisions are made, the model determines the availabilities of the ambulances ( $x_d^j$ ) and ambulance dispatching decisions ( $y_d^j$ ) given a scenario. The objective function (5) maximizes the covered demands in scenario  $\xi^r$ . Constraint (6) ensures that at most one ambulance is dispatched to serve a call. Constraint (7) guarantees that a task to respond to call  $d$  can be assigned to an ambulance station only when the station currently has at least one ambulance.

Constraints (8) and (9) ensure that ambulances are assigned based on a predetermined dispatching policy. Constraint (8) determines the number of available ambulances at location  $j$  when call  $d$  arrives. From the number of ambulances at  $j$  when call  $(d - 1)$  arrived, we subtract the number of ambulances (0 or 1) dispatched from station  $j$  to call  $(d - 1)$ , and add the number of ambulances returning to the station before call  $d$  arrives. Constraint (9) uses the priority set  $L_d^j$  and the availability of ambulances therein to make the dispatching decision,  $y_d^j$ . By constraint (9), incoming call  $d$  is assigned to an available ambulance at station  $j$  if and only if station  $j$  has at least one available ambulance (i.e.,  $x_d^j \geq 1$ ) and there is no available ambulance that has a higher priority for call  $d$  than station  $j$  (i.e.,  $\sum_{i \in L_d^j} x_d^i = 0$ ). For example, suppose that ambulance station  $j$  has the highest priority for incoming call  $d$ , and thus  $L_d^j$  is empty. Then, as long as station  $j$  has at least one available ambulance, constraint (9) requires  $y_d^j = 1$  to assign call  $d$  to station  $j$ .

## 4 Solution methods

In this section, we discuss methods for solving the proposed model. Our model is a stochastic integer program, and it involves integer variables for the first and second stages. Existing solution techniques for stochastic program are designed mostly for problems with continuous variables and deemed ineffective for stochastic integer program due to the integrality restrictions (Birge and Louveaux 2011).

We solve the difficulty by developing an algorithm that utilizes logic-based Benders decomposition. Unlike nominal Benders decomposition, which requires the duality of a sub-problem to generate a cut, logic-based Benders decomposition uses a relatively simple cut. Theoretically the algorithm can produce an optimal solution for the optimization problem in a finite number of steps. Unfortunately, our initial numerical experience indicates that the rate of convergence is too slow for our problem. We addressed this problem by introducing a few approximation schemes for the implementation of the algorithm.

#### 4.1 Proposed approach: logic-based Benders decomposition

A Benders decomposition is a solution framework for solving a large-sized optimization problem. Application of the Benders decomposition involves the structural decomposition of a problem into a master problem and one or more of its sub-problems. The master problem temporarily allows some decision variables to be computed, i.e., to obtain a trial solution, which renders the remaining problems more tractable. In turn, the remaining problems, i.e., the sub-problems, are solved given the trial solution. The sub-problems evaluate the feasibility and optimality of the trial solution and provide this information to the master problem as a cut. If the sub-problems determine the trial solution to be infeasible, a cut to exclude the particular solution is added to the master problem, i.e., a feasibility cut. When the solution is found to be feasible by the sub-problems, a cut is generated to create a tighter bound for the master problem, i.e., an optimality cut. The master problem is then re-solved with the updated cuts. These procedures are repeated until the objective value of the master problem is either sufficiently close to the bounds determined by the sub-problems or the master problem becomes infeasible.

The Benders decomposition was extended by Slyke and Wets (1969). Slyke and Wets (1969) proposed an L-shaped method to solve a stochastic programming model. In stochastic programming, the first and second stage problems correspond to the master and sub-problems of the Benders decomposition, respectively. Here it should be noted that, to generate the cut, the Benders decomposition and L-shaped method use the duality of the sub-problems. As mentioned earlier, our problem is a stochastic integer program, which means the duality of the sub-problems is lost; thus, existing solution methods are difficult to apply.

In terms of addressing the difficulty, Hooker and Yan (1995) proposed a logic-based Benders decomposition, which is structurally similar to a Benders decomposition, except that it uses relatively simple logical expressions, i.e., *no good* cuts, to represent the feasibility and optimality of a trial solution. When  $[x_1, x_2, \dots]$  represents a 0–1 variable for a master problem, an example of a no good cut intended to exclude an infeasible solution  $\star$  is shown below:

$$\sum_{i \in T} x_i - \sum_{i \in F} x_i \leq |T| - 1, \quad (12)$$

where,  $T = \{i | \tilde{x}_i = 1\}$ , and  $F = \{i | \tilde{x}_i = 0\}$  (Jain and Grossmann 2001). The algorithm also uses a logical expression to generate the optimality cut and it generally depends on the problem-specific properties of a target problem.



In this work, we apply a logic-based Benders decomposition to solve our problem (2–11). The approach involves the use of a master problem to determine an initial location solution without considering the actual coverage of the solution. The solution is then evaluated by solving the sub-problems (5–11) for all scenarios, given the solution. The results of the sub-problems are delivered to the master problem as cuts, which provide bounds on the objective values of the location solutions of the master problem. Accordingly, the master problem is re-solved with the updated cuts. These procedures are repeated until the optimal solution of the master problem is equal or close to the bounds determined by the cuts. The master problem in iteration  $K$  has the following form:

$$\begin{aligned}
 (P_K^{master}) \quad & \max \quad \theta \\
 \text{s.t.} \quad & (3), (4)
 \end{aligned}
 \tag{13}$$

$$\theta \leq \sum_{j \in J^k} c_j^k \cdot x_0^j + \sum_{j \in J \setminus J^k} M \cdot x_0^j \quad \forall k \in \{1, \dots, K - 1\},
 \tag{14}$$

$$\theta \leq B(k) \quad \forall k \in \{1, \dots, K - 1\}.
 \tag{15}$$

The master problem determines a solution for the location of ambulances subject to the original constraints (3) and (4), and two types of optimality cuts (14) and (15). We note that the master problem does not contain a feasibility cut because all possible locations for  $p$  ambulances are feasible in our problem; hence, the sub-problems only generate optimality cuts.

Optimality cut (14) sets the objective value of the trial location solutions generated before the current iteration. Let  $J^k$  denote a set of locations for the location solution of the master problem in iteration  $k$ , i.e.,  $\{j \in V | x_0^{j,k} > 0\}$ . Here,  $x_0^{j,k}$  represents the location solution for station  $j$  in iteration  $k$  and  $c_j^k$  is the expected number of covered demands by an ambulance located at station  $j$ . Given an ambulance location solution and a scenario, it is straightforward to assign the available ambulances to incoming calls following a specific dispatching policy. Based on the assignments, we calculate  $c_j^k$  in (14), where  $M$  is a sufficiently large number. By using this notation, (14) provides the actual objective values of the solutions generated until the current iteration.

Cut (15) is the Benders optimality cut generated following Slyke and Wets (1969). Based on LP relaxation and the duality theory, cut (15) provides the upper bound on the  $(P_K^{master})$ . Relaxing the integrality condition on  $x_d^j$  and  $y_d^j$  enables us to generate the Benders optimality cut proposed in Slyke and Wets (1969). Given the master problem solution  $\bar{\mathbf{x}} = [\bar{x}_0^1, \dots, \bar{x}_0^{|J|}]$  at an iteration, the LP-relaxed sub-problem for scenario  $\zeta^r$  is written as follows:

$$\begin{aligned}
 (LP_r^{sub}(\bar{\mathbf{x}})) \quad & \max \quad \sum_{d \in D} \sum_{j \in V} W_d^j \cdot y_d^j \\
 \text{s.t.} \quad & (6) - (9)
 \end{aligned}
 \tag{16}$$

$$x_0^j = \bar{x}_0^j \quad \forall j \in V, \quad (17)$$

$$x_d^j \leq p \quad \forall d \in D^r, j \in V, \quad (18)$$

$$y_d^j \leq 1 \quad \forall d \in D^r, j \in V, \quad (19)$$

$$x_d^j, y_d^j \geq 0 \quad \forall d \in D^r, j \in V. \quad (20)$$

Note that  $(LP_r^{sub}(\mathbf{x}))$  is finite. Then the current optimal dual solution for  $(LP_r^{sub}(\bar{\mathbf{x}}))$  is a feasible solution for the dual of  $(LP_r^{sub}(\mathbf{x}))$  for all  $\mathbf{x}$  which satisfies constraints (3) and (4). In addition, the current dual solution is not necessarily an optimal solution for the dual of  $(LP_r^{sub}(\mathbf{x}))$ . Hence, by the duality theory we have an upper bound on the  $(LP_r^{sub}(\mathbf{x}))$ . We also note that  $(LP_r^{sub}(\bar{\mathbf{x}}))$  is an upper bound of the second stage of problem (5–11) by the LP relaxation. Then based on the conditions, we generate the cut for the next iteration by solving the LP relaxed sub-problems for all scenarios and using the optimal dual multipliers, that is,

$$\theta \leq B(k) = \sum_{r=1}^N p^r \left\{ \sum_{j \in V} \alpha^{j,r} \cdot x_0^j + \sum_{d \in D^r} \sum_{j \in V} (p \cdot \beta_d^{j,r} + \gamma_d^{j,r}) \right\}, \quad (21)$$

where  $\alpha^{j,r}$ ,  $\beta_d^{j,r}$ , and  $\gamma_d^{j,r}$  are dual variables of  $(LP_r^{sub}(\bar{\mathbf{x}}))$  for constraints (17), (18), and (19), respectively. For more details on the Benders cut, we refer the reader to Slyke and Wets (1969).

## 4.2 Implementation

As described in Sect. 4.1, a logic-based Benders decomposition iteratively tightens the gap between the upper and lower bounds on the optimal value, by adding the optimality cuts. Therefore, for problems with a finite number of solutions, a logic-based Benders decomposition theoretically yields an optimal solution of the problem in a finite number of iterations. However, our initial tests show that the algorithm is unable to provide an optimal solution within a practically feasible computation time.

Reduction of the computation time requires the generation of strong cuts that incorporate the structure of the target problem. Unfortunately, the optimality cuts (14) and (15) do not have the required strength for the following reasons. Cut (14) only evaluates a current trial solution and does not provide information for other solutions with similar characteristics to the solution. Moreover, cut (15) provides information for the LP-relaxed objective value of a location solution, rather than for the original problem.

We decided to resolve the difficulty by applying a few approximation schemes to the algorithm. First, we use variable neighborhood search (VNS) to generate an initial trial solution of the master problem. This ensures that the algorithm has a tight lower bound on the optimal value. Next, we terminate the algorithm if the current incumbent solution is not improved after a certain number of iterations.

In VNS, we first define several neighborhood structures,  $\mathcal{N}_k (k = 1, \dots, k_{max})$ . A neighborhood structure specifies the distance between two candidate solutions, which is used to identify neighbors for the current solution  $\mathbf{x}$ . VNS uses several neighborhood structures to avoid local optima by exploring a large solution space, including distant neighborhoods of the current solution. The solution structure for the location problem is simple and allows for easy measurement of the distance between two feasible solutions. For these reasons, the VNS algorithm can readily be implemented for location problems. We follow the basic structure of VNS described in Hansen and Mladenović (2001), as set out in Algorithm 1.

In Algorithm 1, we first define the set of neighborhood structures as follows:

$$\mathcal{N}_k(\mathbf{x}) = \{\mathbf{x}' : |\mathbf{x} \setminus \mathbf{x}'| = |\mathbf{x}' \setminus \mathbf{x}| = k\}.$$

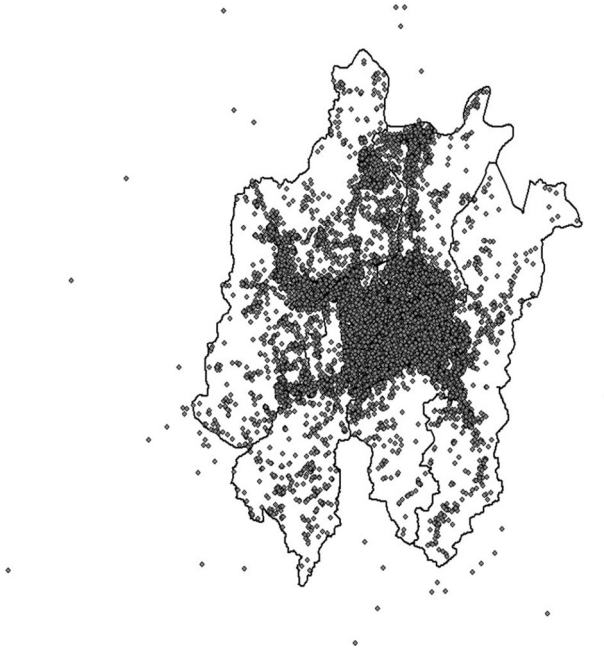
If a location solution  $\mathbf{x}'$  differs from  $\mathbf{x}$  in  $k$  locations, i.e.,  $|\mathbf{x} \setminus \mathbf{x}'| = |\mathbf{x}' \setminus \mathbf{x}| = k$ , then  $\mathbf{x}'$  belongs to a neighborhood of  $\mathbf{x}$  in neighborhood structure  $\mathcal{N}_k$ . *Shaking*( $\mathbf{x}, k$ ) randomly generates a solution  $\mathbf{x}'$  from the  $k^{\text{th}}$  neighborhood of  $\mathbf{x}$ . After the solution  $\mathbf{x}'$  is obtained, a local search method *LocalSearch*( $\mathbf{x}'$ ) is applied to improve solution  $\mathbf{x}'$ . In our implementation, we search all neighborhoods of  $\mathbf{x}'$  in  $\mathcal{N}_2$  and return the best solution among them. Then, the resulting solution  $\mathbf{x}''$  is accepted if  $\mathbf{x}''$  is an improvement of the current incumbent solution.

## 5 Computational experiments

We test our algorithm on problem instances generated from the real EMS call data. We use the EMS call data from Daejeon, a major city in Korea. Daejeon's population is approximately 1.5 million, and the area is 540 km<sup>2</sup>. In 2010, there were total of 59,359 EMS calls in the city. The map of Daejeon with the EMS call locations in the city is shown in Fig. 1.

To obtain ambulance location solutions by our algorithm, we first generate scenarios for the model. We randomly select 50 days from year 2010, and use the actual EMS call arrival data for those days to construct 50 scenarios. So each scenario corresponds to a sequence of actual call arrivals on a particular day. Using the 50 scenarios, our model determines optimal ambulance locations.

For comparison, we also obtain ambulance location solutions by using two classical ambulance location models, a model for the backup coverage problem (BACOP2) (Hogan and ReVelle 1986) and MALP II (ReVelle and Hogan 1989). BACOP2 maximizes the weighted sum of two types of objective functions. The first objective function maximizes the number of demands covered at least once, and the second objective function maximizes the number of demands covered twice, i.e., back-up coverage. MALP II is an extension of MALP, obtained by relaxing an assumption in MALP, i.e., the busy fractions for all ambulances are identical. As a result, the number of ambulances required to cover a demand with a certain reliability level differs for each demand.



**Fig. 1** Demand nodes in Daejeon 2010

In the experiments, we arbitrarily chose 30 candidate sites for ambulance locations. For a dispatching policy, we apply the nearest available policy. 8 min is used as the time standard for successful response; if an EMS call is responded (i.e., an ambulance arrives at the call location) within 8 min, it is considered a successful response. Using this setting, we implement the three location models in Java, and used CPLEX v.12.5 to solve the models. We varied the number of ambulances  $p$ , and compared the performance from the location solutions by the three approaches.

Evaluations of the obtained location solutions are done by a simple, deterministic simulation as follows. For each evaluation of a location solution, we first set the number of ambulances at each station as given by the location solution. Then we sequentially generate calls at a specific time and location according to the actual data for the entire year. For each arriving call, we use a specified dispatching policy to select an ambulance to send to the call. Then for each response, we record its response time by computing the travel time by the ambulance to the call location. After a predetermined turnaround time, the dispatched ambulance returns to its home station and becomes available for a next service task. Finally, the simulation model reports the percentage of covered demands based on the records.

Table 1 lists the results for  $p = 8, 10, 15,$  and  $20$ . Note that solving BACOP2 and MALP II requires some model parameter values: weighting factor between single and back-up coverage objectives in BACOP2, and service reliability level in MALP II, respectively. Instead of arbitrarily setting these parameters, we solve BACOP2 and MALP II multiple times by varying their values, and report the best results

**Table 1** Percentage (%) of covered demands

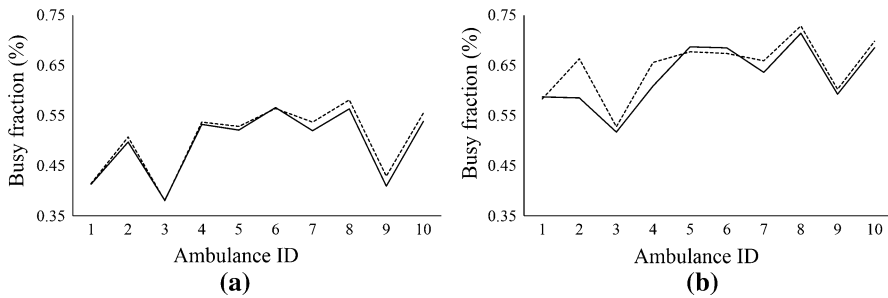
$p$	Proposed	BACOP2	MALP II
8	61.8	44.3	57.6
10	73.5	57.3	66.7
15	89.2	76.6	86.0
20	94.6	84.4	93.9

obtained. For BACOP2, we increase the weights from 0 to 1 by an increment of 0.1. For MALP II, we use 85, 90, and 95% for the reliability level following ReVelle and Hogan (1989).

In Table 1, we report the percentage of successful responses by solutions from the three approaches. The results indicate that our algorithm outperforms the other location models. BACOP2 model performs far worse than the other two models. BACOP2 model tends to spread out ambulances to maximize the deterministic coverage. But its deterministic approach, even with the back-up coverage requirement, does not effectively incorporate the availability issue of ambulances. On the other hand, MALP II model tends to collocate ambulances due to the reliability constraint in the model. This collocation strategy may not be very effective particularly when there are only a small number of ambulances; it increases the level of service for possibly small volume of demand at the cost of very low level of service for larger areas. Table 1 shows that the gap between MALP II and the proposed model is slightly larger when the number of ambulances  $p$  is smaller.

## 6 Discussion

The performance gain achieved by our model is attributed by the fact that our model explicitly computes the availability of ambulances in computing location solutions. It incorporates two key factors into the ambulance location model: temporal variations in call arrivals and a dispatching policy. Most probabilistic location models handles the availability of ambulances by using its busy fraction, and appropriately estimating busy fractions is a fundamental challenge (ReVelle and Hogan 1989). Furthermore, they typically model it as a constant with respect to the time of day, as in a homogeneous Poisson process. However, there clearly exists a temporal variation in call arrivals throughout a day. More calls arrive in the late AM and early PM, and fewer calls arrive in the middle of the night. This temporal variation makes the estimation of the busy fraction deviate further from the true value. It is known that the availability of ambulances is also affected by a dispatching policy as well, and there are several studies that develop optimal dispatching policies (Aboueljineane et al. 2013; Andersson and Värbrand 2007; Jagtenberg et al. 2016; Lee 2011; Lim et al. 2011). As will be shown in this section, dispatching policies change optimal locations of ambulances, and an inappropriate assumption about dispatching policy can negatively affect the performance of the location solution.



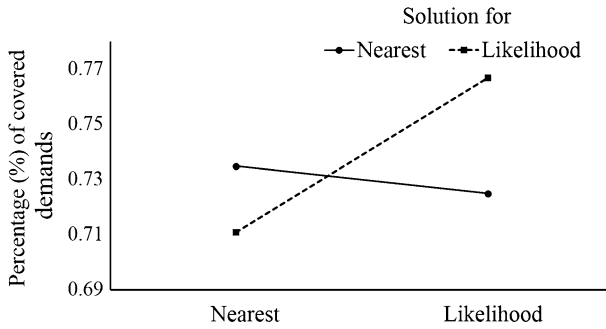
**Fig. 2** Busy fraction of ambulances from our model (*solid*) and from one-year operation (*dotted*) for 10-ambulance case **a** average busy fraction, **b** busy fraction for peak hours (9 AM–12 PM)

Let us first examine the busy fraction of ambulances computed in our model. Since our model makes dispatching decisions and keeps track of availability of ambulances at each station, we can compute the busy fraction of ambulances from the model. Specifically, we have  $y_d^j$  that describes the history of ambulance operation from which we can compute the busy fraction for ambulances. Note that since we use sampled call data to generate call arrival scenarios, the computed busy fraction is an approximation for the “actual” busy fraction. The actual busy fraction can be easily computed by the same simulation described in Sect. 5 using entire call data of 2010.

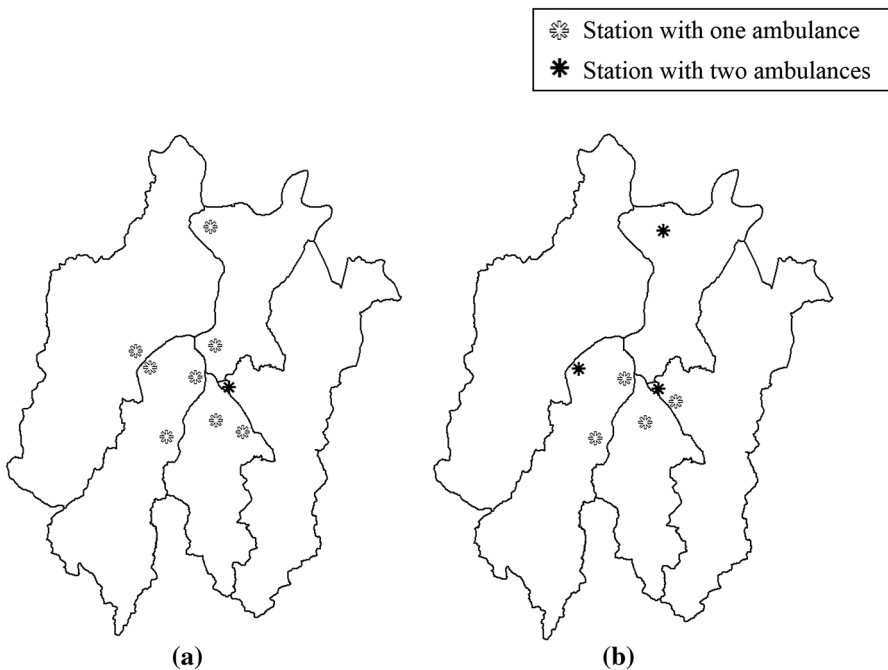
Figure 2 shows the busy fraction values computed from the model (solid line) versus from the actual 1-year operation (dotted line). Results shown in Fig. 2 clearly demonstrate that the busy fraction in our location model does match the actual busy fraction from one-year of operation under the obtained solution. Recall that it is a typical problem in a probabilistic location model that the busy fraction assumed in a location model is indeed different from the resulting busy fraction when the obtained location solution is used. Figure 2 suggests that our model determines a location solution based on the accurately represented availability of ambulances.<sup>2</sup> In addition, our model adequately captures the dynamics of the ambulance operation due to the temporal variations in call arrivals. As shown in Fig. 2b, busy fraction during the peak hours is quite higher than the average value in Fig. 2a, and it closely follows the actual busy fraction as well.

Next, we examine the effects of the dispatching policy on the ambulance location solution. As we mentioned earlier, there exists an interaction between a dispatching policy and a location solution of ambulances. We investigate this interaction by deriving ambulance location solutions under different dispatching policy. For this test, we use the least likelihood dispatching policy proposed in Repede and Bernardo (1994). The least likelihood dispatching policy works as follows; it sends the nearest available ambulance to an arriving call when that ambulance can arrive at the call location within the time standard. Otherwise it dispatches an ambulance that has the least likelihood of receiving a call. In the experiments, location

<sup>2</sup> Note that unlike a probabilistic location model our model does not utilize the busy fraction to compute a location solution. They are only implicit in the model.



**Fig. 3** Interaction effect between a dispatching policy and an ambulance location solution



**Fig. 4** Map of Daejeon with location solutions for **a** nearest available dispatching policy, **b** least likelihood dispatching policy

solutions for each dispatching policy when  $p = 10$  are derived, and these solutions are evaluated by a simulation model under the two different dispatching policies. The results are presented in Fig. 3.

In Fig. 3, x-axis is a dispatching policy used in the simulation model. Then simulated results (i.e., the percentage of covered demands) for each location solution obtained under the nearest available dispatching policy (solid line) and the least likelihood dispatching policy (dotted line), are plotted on y-axis. From Fig. 3, we first see the interaction effect between a dispatching policy and an ambulance location solution. A solution derived under one dispatching policy is not the best

solution when the other policy is in fact used for actual operation. The performance of an ambulance location solution is quite sensitive to a dispatching policy. For the location solution under the least likelihood dispatching policy, the percentage of covered demands is decreased by almost 6% when the other policy was used in the operation. As such, the performance of a location solution critically depends on which dispatching policy is used in the actual operation of ambulances.

Configurations of the ambulance locations for each dispatching policy are also quite different. Figure 4 shows the ambulance location solutions obtained under the two dispatching policies. Under the least likelihood dispatching policy, more ambulances are collocated at some stations whereas in the nearest available dispatching policy, the location solution tends to locate a single ambulance at a station. These results verify that considering a dispatching policy is crucial for determining ambulance locations.

## 7 Conclusion

In this paper, we propose an ambulance location model designed to explicitly compute the availability of ambulances while considering the two key factors in ambulance location decisions: temporal variations in call arrivals and a dispatching policy. The model is based on a stochastic programming approach and the temporal variations in call arrivals are represented as a set of scenarios. Moreover we explicitly incorporate the ambulance dispatching policy into the ambulance location problem such that the interaction between two decisions, i.e., ambulance dispatching and locations, is considered. It allows us to incorporate ambulance availability in a more precise fashion than classical probabilistic location models. The proposed model was solved by developing an algorithm based on a logic-based Benders decomposition. The algorithm theoretically guarantees the optimality of the solution. However, as the algorithm required a significant amount of time to obtain an optimal location solution, we implemented a few approximation schemes using a meta-heuristic algorithm. The experiments demonstrate that the proposed model outperforms some of the classic location models.

Let us conclude this paper by discussing potential difficulties in implementing the proposed algorithm. When the number of demand sites increases or when a random vector for call arrivals has a large number of possible realizations, the number of scenarios required to properly represent the call arrivals is likely to increase substantially. It results in excessive computational burden, making the scenario-based approach potentially intractable. This is one of the primary challenges in stochastic programming. To resolve this issue, we need research to develop techniques to effectively generate scenarios within the allowed computational budget and to evaluate the goodness of the generated scenarios.

**Acknowledgements** This research was supported by a Grant “Research and development of modeling and simulating the rescues, the transfer, and the treatment of disaster victims” [MPSS-SD-2013-36] through the Disaster and Safety Management Institute funded by Ministry of Public Safety and Security of Korean government.



## References

- Aboueljainane L, Sahin E, Jemai Z (2013) A review on simulation models applied to emergency medical service operations. *Comput Ind Eng* 66(4):734–750
- Andersson T, Värbrand P (2007) Decision support tools for ambulance dispatch and relocation. *J Oper Res Soc* 58(2):195–201
- Beraldi P, Bruni M (2009) A probabilistic model applied to emergency service vehicle location. *Eur J Oper Res* 196(1):323–331
- Birge JR, Louveaux F (2011) Introduction to stochastic programming. Springer, New York
- Brotcorne L, Laporte G, Semet F (2003) Ambulance location and relocation models. *Eur J Oper Res* 147(3):451–463
- Carson YM, Batta R (1990) Locating an ambulance on the Amherst Campus of the State University of New York at Buffalo. *Interfaces* 20(5):43–49
- Church R, Velle CR (1974) The maximal covering location problem. *Pap Reg Sci* 32(1):101–118
- Daskin MS, Dean LK (2005) Location of health care facilities. In: Sainfort F, Brandeau M, Pierskalla W (eds) *Operations research and health care*. Springer, New York, p 43–76
- Farahani RZ, Asgari N, Heidari N, Hosseini M, Goh M (2012) Covering problems in facility location: a review. *Comput Ind Eng* 62(1):368–407
- Hakimi SL (1964) Optimum locations of switching centers and the absolute centers and medians of a graph. *Oper Res* 12(3):450–459
- Hansen P, Mladenović N (2001) Variable neighborhood search: principles and applications. *Eur J Oper Res* 130(3):449–467
- Hogan K, ReVelle C (1986) Concepts and applications of backup coverage. *Manag Sci* 32(11):1434–1444
- Hooker JN, Yan H (1995) Logic circuit verification by benders decomposition. *Principles and practice of constraint programming: the newport papers*, pp 267–288
- Huang Y, Fan Y (2011) Modeling uncertainties in emergency service resource allocation. *J Infrastruct Syst* 17(1):35–41
- Jagtenberg C, Bhulai S, Van der Mei R (2016) Dynamic ambulance dispatching: is the closest-idle policy always optimal? *Health Care Manag Sci*. doi:10.1007/s10729-016-9368-0
- Jain V, Grossmann IE (2001) Algorithms for hybrid MILP CLP models for a class of optimization problems. *INFORMS J Comput* 13(4):258–276
- Jia H, Ordóñez F, Dessouky M (2007) A modeling framework for facility location of medical services for large-scale emergencies. *IIE Trans* 39(1):41–55
- Lee S (2011) The role of preparedness in ambulance dispatching. *J Oper Res Soc* 62(10):1888–1897
- Li X, Zhao Z, Zhu X, Wyatt T (2011) Covering models and optimization techniques for emergency response facility location and planning: a review. *Math Methods Oper Res* 74(3):281–310
- Lim CS, Mamat R, Bräunl T (2011) Impact of ambulance dispatch policies on performance of emergency medical services. *IEEE Trans Intell Transp Syst* 12(2):624–632
- Marianov V, ReVelle C (1996) The Queueing Maximal availability location problem: a model for the siting of emergency vehicles. *Eur J Oper Res* 93(1):110–120
- Matteson DS, McLean MW, Woodard DB, Henderson SG (2011) Forecasting emergency medical service call arrival rates. *Ann Appl Stat* 5(2B):1379–1406
- Nickel S, Reuter-Oppermann M, Saldanha-da Gama F (2016) Ambulance location under stochastic demand: a sampling approach. *Oper Res Health Care* 8:24–32
- Owen SH, Daskin MS (1998) Strategic facility location: a review. *Eur J Oper Res* 111(3):423–447
- Pereira MA, Coelho LC, Lorena LA, De Souza LC (2015) A hybrid method for the probabilistic maximal covering location-allocation problem. *Comput Oper Res* 57:51–59
- Repede J, Bernardo JJ (1994) Case study developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky. *Eur J Oper Res* 75:567–581
- ReVelle C, Eiselt H (2005) Location analysis: a synthesis and survey. *Eur J Oper Res* 165(1):1–19
- ReVelle C, Hogan K (1989) The maximum availability location problem. *Transp Sci* 23(3):192–200
- Schilling DA (1982) Strategic facility planning: the analysis of options. *Decis Sci* 13(1):1–14

- Toregas C, Swain R, ReVelle C, Bergman L (1971) The location of emergency service facilities. *Oper Res* 19(6):1363–1373
- Van Slyke RM, Wets R (1969) L-shaped linear programs with applications to optimal control and stochastic programming. *SIAM J Appl Math* 17(4):638–663

**Inkyung Sung** is a research associate in the Department of Industrial and Systems Engineering at KAIST. His research interests focus on optimization and scheduling for healthcare service delivery systems and disaster management. He received the Ph.D. degree in Industrial and Systems Engineering from KAIST.

**Taesik Lee** is an associate professor of Industrial and Systems Engineering Department at KAIST, Korea. His research interests are primarily in the area of mathematical modeling and system simulation, especially for healthcare service delivery system. For the past few years, he has been focusing on various aspects of emergency medical service systems in the context of disaster response management. Dr. Lee obtained his MS and Ph.D. degree from MIT in the Department of Mechanical Engineering.