Ambulance Location Problem with Stochastic Call Arrivals Under Nearest Available Dispatching Policy

Inkyung Sung and Taesik Lee

Introduction

An ambulance location problem has been extensively studied since as early as 1970s (Church and ReVelle 1974; Toregas et al. 1971). The problem is to determine the locations of ambulances to provide maximum coverage to potential demand sites. Locations of available ambulances are a major factor to determine the response times to arriving calls. Brotcorne et al. (2003); Farahani et al. (2012); Li et al. (2011); Owen and Daskin (1998); ReVelle and Eiselt (2005) provide a comprehensive review of location problems in emergency medical service (EMS) systems.

Ambulance location problems are often formulated as a covering problem. A demand site is considered covered if it can be reached from an ambulance station within a time standard. Then, the problem finds optimal number and locations for ambulances so that the sum of covered demand sites is maximized. The classic ambulance location problems model the coverage as deterministic. It assumes that ambulances are always available to respond to emergency calls. On the other hand, more recent models incorporate randomness in ambulance's availability. These models often use the concept of busy fraction of an ambulance—probability for being unavailable to respond to a call.

Our study has been motivated by the fact that an ambulance dispatching policy is an important factor affecting ambulance's availability. An ambulance dispatching policy determines which of the ambulances available at the moment is sent to serve an incoming call. A choice made for the current call determines the available ambulances and their coverage for the next arriving call. This implies that there

I. Sung • T. Lee (✉)
KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 305-701, Republic of Korea
e-mail: inkyung@kaist.ac.kr; taesik.lee@kaist.edu

A. Matta et al. (eds.), *Health Care Systems Engineering for Scientists and Practitioners*, Springer Proceedings in Mathematics & Statistics 169,
DOI 10.1007/978-3-319-35132-2_10

is an interaction effect between the location decision and dispatching policy; an optimal location solution under one dispatching policy may not be the optimal solution for another policy. Therefore ambulance locations should be determined while considering an ambulance dispatching policy. We construct our model for a specific dispatching policy to explicitly incorporate the effect of a dispatching policy on ambulance location solution. Given a temporal sequence of call arrivals, which is sampled from real data, the model simultaneously determines ambulance locations and call assignments to ambulances under the dispatching policy.

Our model is also designed to take into account variations in call arrivals. Most of the probabilistic models in the ambulance location literature assume stationary call arrivals and use an average call arrival rate in the model. However, the actual EMS call data that the volume of call arrivals varies significantly over the course of a day, weekdays vs. weekends, and between seasons (e.g., Matteson et al. 2011). To properly represent the variations in call arrivals, we take a stochastic programming approach and incorporate the uncertainty of call arrivals into location decisions.

Problem Formulations and Scenario Decomposition Implementation

We develop a solution algorithm for an ambulance location problem that takes into account an ambulance dispatching policy and variations in EMS call arrivals. For this, we apply a stochastic programming approach. Stochastic programming is a framework for modeling an optimization problem with two types of decision variables, *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 the here-and-now decision made earlier is adapted accordingly. In general, a recourse decision depends upon a here-and-now decision. With the decision structure, stochastic programming enables to consider uncertainties in the outcomes and derive a solution that is robust to the uncertainties. Stochastic programming uses a set of scenarios, i.e., possible futures, and derives solutions that perform well across all scenarios. Scenarios is a realization of the uncertainties.

In our algorithm, an ambulance location decision corresponds to the here-and-now decision, and a dispatching is the recourse decision. A scenario in our problem is defined by a sequence of call arrivals. We determine ambulance locations before observing call arrivals under a particular scenario. Then after observing actual call arrivals, ambulance dispatching decisions are made given the ambulance location decision. By taking a stochastic programming approach, we derive an ambulance location solution that performs well across all possible scenarios.

Scenario Decomposition for Stochastic Programming

Let $\boldsymbol{\xi}$ denote a random vector for call arrivals with a support Ξ and known distribution P. We assume that $\boldsymbol{\xi}$ has a finite support, and there are N realizations $\boldsymbol{\xi}^r$, $r \in \{1, \ldots, N\}$. A realization of the random vector is referred to as a *scenario*. We consider a stochastic program for an ambulance location problem:

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

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

One of the approaches to solve (1) is *scenario decomposition*. The main idea of scenario decomposition is to decompose the main problem (1) into N sub-problems by maintaining an individual copy of here-and-now decision variable for each scenario, $\mathbf{x}^r$. Because location solutions should not depend on a scenario, *non-anticipativity* constraint is imposed to require $\mathbf{x}^1 = \cdots = \mathbf{x}^N$. With the non-anticipativity constraint represented by the equality $\sum_{r=1}^{N} \mathbf{A}^r\mathbf{x}^r = 0$, (1) can be rewritten as

$$\max \left\{ \sum_{r=1}^{N} f_r(\mathbf{x}^r) : \mathbf{x}^r \in X \;\; \forall r, \;\; \sum_{r=1}^{N} \mathbf{A}^r\mathbf{x}^r = 0 \right\}, \tag{2}$$

where $f_r(\mathbf{x}^r) = p_r \cdot f(\mathbf{x}^r, \boldsymbol{\xi}^r)$. Unfortunately, it is difficult to incorporate the non-anticipativity constraint to the sub-problems. To overcome the difficulty, Lagrangian relaxation is applied to (2) (Fisher 2004). By dualizing the non-anticipativity constraint, the Lagrangian dual of (2) is obtained as follows:

$$\min_{\lambda} \left\{ z_D = \sum_{r=1}^{N} \max\{f_r(\mathbf{x}^r) + \boldsymbol{\lambda}\mathbf{A}^r\mathbf{x}^r : \mathbf{x}^r \in X \;\; \forall r\} \right\}, \tag{3}$$

where $\boldsymbol{\lambda}$ is a dual vector. This provides an upper bound for (2). By choosing $\boldsymbol{\lambda}$ such that $\mathbf{x}^r$ is identical for all $r \in \{1, \ldots, N\}$, we can find an optimal solution of (2) (Carøe and Schultz 1999). To solve (3), we implement a simple algorithm, following an algorithm proposed by Ahmed (2013). The algorithm is shown in Algorithm 1.

Algorithm 1 produces candidate solutions of the sub-problems for each scenario and calculates upper bound by the sum of the Lagrangian objective functions of the candidate solutions. A lower bound is also calculated by evaluating the original objective values of the candidate solutions. After obtaining the candidate solutions for each scenario, Algorithm 1 updates $\boldsymbol{\lambda}$ such that the upper bound is tightened. Then this procedure is repeated until the gap between lower and upper bound is close enough or the number of iterations reaches a certain threshold value.

Algorithm 1 Scenario decomposition

```
$UB \leftarrow \infty$, $LB \leftarrow -\infty$, $\mathbf{x}^* \leftarrow \emptyset$, $n_I \leftarrow 0$, $\boldsymbol{\lambda} \leftarrow \mathbf{0}$
while $UB - LB > \varepsilon$ and $n_I < n_{max}$ do
    $n_I \leftarrow n_I + 1$
    for $r = 1$ to $N$ do
        solve $\max\{f_r(\mathbf{x}) + \boldsymbol{\lambda}\mathbf{A}^r\mathbf{x}\}$
        let $v^r$ be the optimal value and $\mathbf{x}^r$ be an optimal solution
        $u \leftarrow 0$
        for $r' = 1$ to $N$ do
            compute $f_{r'}(\mathbf{x}^r)$ and set $u \leftarrow u + f_{r'}(\mathbf{x}^r)$
        end for
        if $LB < u$ then
            $LB \leftarrow u, \mathbf{x}^* \leftarrow \mathbf{x}^r$
        end if
    end for
    $UB \leftarrow \sum_{r=1}^{N} v^r$
    update $\boldsymbol{\lambda}$
end while
```

To update $\boldsymbol{\lambda}$, we use a sub-gradient method. Given $\boldsymbol{\lambda}'$, the sub-gradient method calculates the gradient of z_D at $\boldsymbol{\lambda}'$. The gradient at $\boldsymbol{\lambda}'$ is given by $\sum_{r=1}^{N} \mathbf{A}^r\mathbf{x}^r$, where $\mathbf{x}^r$ is the solution to $\max\{f_r(\mathbf{x}) + \boldsymbol{\lambda}'\mathbf{A}^r\mathbf{x}\}$. Then the method updates the current $\boldsymbol{\lambda}'$ by using the gradient such that (3) can be minimized.

Modeling for the Ambulance Location Problem Given a Single Scenario

In the solution approach described in section "Scenario Decomposition for Stochastic Programming", we need to solve for each scenario the following problem:

$$\max\{f_r(\mathbf{x}) + \boldsymbol{\lambda}\mathbf{A}^r\mathbf{x}\} \tag{4}$$

We formulate this problem as an integer program. Recall that we want to incorporate the effect from a dispatching policy into the location decision. This is achieved by introducing a constraint to ensure that ambulances are assigned to arriving calls based on a chosen dispatching policy. In a sense, it is a location-routing problem with a restriction that a routing is determined once a location decision is given. For our problem, we choose the "nearest available" dispatching policy, which sends the closest available ambulance to an arriving call. This policy is a common practice found in many EMS call centers in Korea.

In the model, we have one principal decision variable and two auxiliary variables:

- integer variable x_t^j that indicates the number of available ambulances at station j at time t

- (auxiliary) binary variable y_d^j that indicates whether a call d is serviced by ambulance located at station j
- (auxiliary) binary variable z_t^j that indicates whether there is any ambulance available at station j right after an ambulance is dispatched at time t

Note that x_0^j defines the ambulance locations at the beginning of a planning horizon, hence it is the location solution for our problem. Also note that y_d^j specifies ambulance dispatching decisions for call d, and it is determined by the nearest available dispatching policy assumed in our model.

Before presenting the full formulation, the notation used in the model is summarized in Table 1.

Here, we rewrite (4) by using W_d^j and a^r. The first term in (4) is the number of covered demands. W_d^j specifies whether demand location d can be covered by station j or not. Using W_d^j, $f_r(\mathbf{x})$ can be written as $\sum_{d\in D}\sum_{j\in V} W_d^j \cdot y_d^j$. The second term in (4) is related to the non-anticipativity constraint. We set the constraint as $(N-1)\mathbf{x}^1 = \mathbf{x}^2+\cdots+\mathbf{x}^N$, and the second term in (4) can be written as $\lambda^j a^r \sum_{j\in V} x_0^j$, where λ^j is jth element of $|V|$ dimensional vector $\boldsymbol{\lambda}$, and $a^r = N-1$ if $r = 1$, otherwise, -1. Then, for given $\boldsymbol{\lambda}$, the objective function can be written as

Table 1 Summary of notation

Symbol	Definition
i, j	Ambulance station index
t	Time index
d	Call index
V	Set of candidate ambulance locations
T	Final time horizon
D	Set of calls occurred during a planning horizon
q	The number of maximum ambulances
a_d	Arrival time of call d
N_t	The number of calls during time interval t
τ_d^j	Distance, measured in time, between call d and station j
R_d^j	Turn around time of call d by an ambulance in station j
S	Time standard for coverage
M	Sufficiently large number
A_t	Set of calls arriving at time t
B_t^j	Set of calls for station j satisfying the condition, $a_d + R_d^j = t$
W_d^j	Constant: 1 if $\tau_d^j \le S$; 0, otherwise
x_t^j	Variable: the number of available ambulances at station j at time t
y_d^j	Variable: 1 if ambulance j dispatched to call d; 0, otherwise
z_t^j	Variable: 1 if at least one ambulance is available at station j right after time t; 0, otherwise

$$\max \sum_{d \in D} \sum_{j \in V} W_d^j \cdot y_d^j + \lambda^r a^r \sum_{j \in V} x_0^j, \tag{5}$$

The constraints of the model are constructed to impose the nearest–available dispatching policy. Before describing the constraints we introduce two assumptions. First, ambulances on its way back to its home station are not available for service until it returns to the station. Second, if a call arrives and all ambulances are busy at the moment, the call is either lost or served from other EMS systems. These assumptions can be justified by the fact that probabilities of the events are very small. In addition, standard practice for operating ambulances in Korea is to return its home station after serving a call in order to get ready for next call arrivals. With these assumptions, the objective function (5) is solved subject to the following set of constraints:

$$\sum_{j \in V} x_0^j = q \tag{6}$$

$$\sum_{j \in V} y_d^j \leq 1 \qquad \forall d \in D \tag{7}$$

$$x_t^j = x_{t-1}^j - \sum_{d \in A_{t-1}} y_d^j + \sum_{d \in B_t^j} y_d^j \qquad \forall t \geq 1, j \in V \tag{8}$$

$$\sum_{d \in A_t} \sum_{j \in V} y_d^j = \min\left(N_t, \sum_{j \in V} x_t^j\right) \qquad \forall t \leq T \tag{9}$$

$$x_t^j - \sum_{d \in A_t} y_d^j \geq z_t^j \qquad \forall t \leq T, j \in V \tag{10}$$

$$x_t^j - \sum_{d \in A_t} y_d^j \leq M \cdot z_t^j \qquad \forall t \leq T, j \in V \tag{11}$$

$$y_d^j \cdot \tau_d^j \leq M(1 - z_t^i) + z_t^i \cdot \tau_d^j \qquad t = a_d, \forall d \in D,\ i,j \in V, i \neq j \tag{12}$$

$$x_t^j \in \mathbb{Z}^+ \qquad \forall t \leq T, j \in V \tag{13}$$

$$y_d^j \in \{0, 1\} \qquad \forall d \in D, j \in V \tag{14}$$

$$z_t^j \in \{0, 1\} \qquad \forall t \leq T, j \in V \tag{15}$$

Constraint (6) limits the total number of ambulances to be located at q. Constraint (7) ensures that at most one ambulance is dispatched to serve a call. Constraint (8) determines the number of available ambulances at station j at the beginning of time interval t, x_t^j. It is computed by subtracting ambulances dispatched from station j to calls during time period $t - 1$ and adding ambulances that are returning to the station at the beginning of time interval t. Constraint (9) ensures the number of ambulances dispatched during time interval t equals to either the number of calls during time interval t or the number of available ambulances at the beginning of the time interval. By constraint (10) and (11), z_t^j becomes one if station j has at least one ambulance at time $t + \epsilon$, where $t + \epsilon$ is the time right after a dispatching decision at station j is made. Then, constraint (12) ensures ambulances are assigned based on the nearest available dispatching policy. It states that travel time for call d from station j is smaller than any other station at which ambulances are available. It should be noted that other types of dispatching policy based on priority (e.g., regionalized response Aboueljinane et al. 2013) can be modeled in the same way by changing the travel time to corresponding cost measures.

In Algorithm 1, we need to solve this sub-problem for all scenarios per each iteration, for a large number of iterations. Therefore it is important to quickly solve the sub-problems to make Algorithm 1 computationally efficient. Unfortunately, our initial tests show that commercial LP solvers—we used CPLEX—cannot solve the sub-problems fast enough, and we decide to develop a meta-heuristic algorithm to quickly obtain near optimal solutions. In this study, we use Variable Neighborhood Search (VNS).

In VNS, we first define several neighborhood structures, $\mathscr{N}_k (k = 1, \ldots, k_{max})$. A neighborhood structure specifies distance between two candidate solutions, which is used to identify neighbors for current solution $\mathbf{x}$. VNS uses several neighborhood structures to avoid local optima by exploring a large solution space, including distant neighborhood of a current solution. The solution structure for location problem is simple and easy to measure the distance between two feasible solutions. For these reasons, VNS algorithm can be readily implemented for location problems. We follow the basic structure of VNS described in Hansen and Mladenovic (2001), and it is depicted in Algorithm 2.

In Algorithm 2, $Shaking(\mathbf{x}, k)$ randomly generates a solution $\mathbf{x}'$ from the k^{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 $\mathscr{N}_2$ and return the best solution among them. Then, the resulting solution $\mathbf{x}''$ is accepted if $\mathbf{x}''$ is better than current incumbent solution.

Algorithm 2 Basic VNS

```
Select the set of neighborhood structures $\mathscr{N}_k$, $k = 1, \ldots, k_{max}$
Generate initial solution, $\mathbf{x}$
repeat
    $k \leftarrow 1$
    repeat
        $\mathbf{x}' \leftarrow Shaking(\mathbf{x}, k)$
        $\mathbf{x}'' \leftarrow LocalSearch(\mathbf{x}')$
        if $accept(\mathbf{x}'')$ then
            $\mathbf{x} \leftarrow \mathbf{x}''$
            $k \leftarrow 1$
        else
            $k \leftarrow k + 1$
        end if
    until $k = k_{max}$
until stopping condition is met
```

For the implementation of VNS, we 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 $\mathcal{N}_k$.

For an initial solution to feed to the VNS algorithm, we solve the integer program (5) without considering the constraints for the nearest available dispatching policy, (9)–(12), (15). That is, we solve for the objective function (5) with a partial set of constraints (6)–(8), (13)–(14). In doing so, we need a new constraint to replace the original constraint (9):

$$\sum_{d \in A_t} y_d^j \leq x_t^j \qquad \forall t \leq T, j \in V.$$

This is to ensure the number of ambulances dispatched during time interval t do not exceed the number of available ambulances available at the beginning of the interval t.

Results

We test our solution algorithm by using EMS call data for the city of Daejeon in Korea. EMS log data for the month of January of 2010 is used to generate a set of scenarios. The algorithm determines ambulance locations based on the scenarios. Then the solution is evaluated by using the call data from February of 2010. For evaluation, we measure the percentage of the calls to which an ambulance arrives within 10 min. For call arrivals, we used the actual data and ambulances are dispatched following the nearest available dispatching policy. As a comparison, we obtain ambulance location solutions by using MALP II (ReVelle and Hogan 1989) and BACOP2 (Hogan and ReVelle 1986). MALP II is a probabilistic ambulance location model to maximize the number of demands covered. Incorporated in the model is the availability constraint, which requires a demand point be covered by multiple number of ambulances. Workload for ambulances determine how many ambulances should cover a demand point to ensure certain level of ambulance availability (60 % in this experiment). BACOP2 also aims to address the availability of ambulances, but it does so by requiring a fixed number of ambulances cover a demand point.

In the experiment, we vary the number of ambulances q to locate, and compare the performance of location solutions by the three approaches. Figure 1 shows the results for $q = 3, 5, 7, 10$.

In Fig. 1, we observe that our algorithm performs better than the other location models. In comparison with MALP II model, when the number of ambulances is

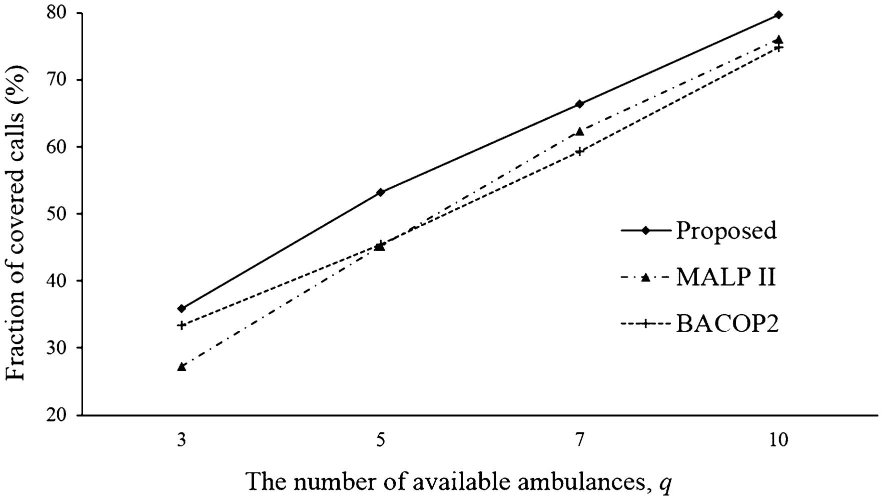

Fig. 1 Fraction of calls served by an ambulance within 10 min

low, the improvements by the proposed algorithm is significant. MALP II model tends to collocate ambulances in order to satisfy the availability constraint, leaving a large number of demands left uncovered. More importantly, it should be noted that MALP II uses busy fraction, which is estimated as an average value and ignores temporal variations in call arrivals. This possibly makes the estimated busy fraction an over-estimation for the night hours and under-estimation for the day hours. The performance gap between the proposed algorithm and BACOP2 increases as the number of ambulances is high. BACOP2 model tends to spread out ambulances in order to maximize the deterministic coverage. While such strategy seems appropriate when the number of ambulances is low, an approach that takes into account the availability of ambulances (the proposed approach and MALP II in this case) offers a larger benefit as the number of ambulances increases.

Conclusion

In this paper, we develop a model and a solution algorithm to locate ambulances. In particular, our model addresses two key factors in ambulance location decisions: a dispatching policy and temporal variations in call arrivals. The novelties of our model are (1) it explicitly describes an ambulance dispatching policy in an ambulance location problem so that the interaction between two decisions, i.e., ambulance dispatching and locations, is considered (2) our model allows us to consider the temporal variations in call arrivals which allows to incorporate ambulance availability in a more precise fashion than classical probabilistic location models.

Specifically, we model this ambulance location problem as an integer program with the constraints for the nearest available dispatching policy. We applied stochastic programming to incorporate various call arrival patterns. To obtain solutions, we implement a scenario decomposition approach which separately solves sub-problems for each scenario while maintaining the non-anticipativity. The solutions of the sub-problems are obtained by using a VNS method, a meta-heuristic algorithm. The experiments demonstrate that by considering the ambulance dispatching policy and temporal variations of call arrivals, it delivers superior performance compared with some of the classic location models.

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" (nema-md-2013-36) from the man-made disaster prevention research center, Ministry of Public Safety and Security.

References

Aboueljinane, L., Sahin E., Jemai Z.: A review on simulation models applied to emergency medical service operations. Comput. Ind. Eng. **66**, 734–750 (2013)

Ahmed, S.: A scenario decomposition algorithm for 0–1 stochastic programs. Oper. Res. Lett. **41**, 565–569 (2013)

Brotcorne, L., Laporte G., Semet, F.: Ambulance location and relocation models. Eur. J. Oper. Res. **147**, 451–463 (2003)

Carøe, C.C., Schultz, R.: Dual decomposition in stochastic integer programming. Oper. Res. Lett. **24**, 37–45 (1999)

Church, R., ReVelle, C.: The maximal covering location problem. Pap. Reg. Sci. **32**, 101–118 (1974)

Farahani, R.Z., Asgari, N., Heidari, N., Hosseininia, M., Goh, M.: Covering problems in facility location: a review. Comput. Ind. Eng. **62**, 368–407 (2012)

Fisher, M.L.: The lagrangian relaxation method for solving integer programming problems. Manag. Sci. **50**, 1861–1871 (2004)

Hansen, P., Mladenovic, N.: Variable neighborhood search: principles and applications. Eur. J. Oper. Res. **130**, 449–467 (2001)

Hogan, K., ReVelle, C.: Concepts and applications of backup coverage. Manag. Sci. **32**, 1434–1444 (1986)

Li, X., Zhao, Z., Zhu, X., Wyatt, T.: Covering models and optimization techniques for emergency response facility location and planning: a review. Math. Methods Oper. Res. **74**, 281–310 (2011)

Matteson, D.S., McLean, M.W., Woodard, D.B., Henderson, S.G.: Forecasting emergency medical service call arrival rates. Ann. Appl. Stat. **5**, 1379–1406 (2011)

Owen, S.H., Daskin, M.S.: Strategic facility location: a review. Eur. J. Oper. Res. **111**, 423–447 (1998)

ReVelle, C.S., Eiselt, H.A: Location analysis: a synthesis and survey. Eur. J. Oper. Res. **165**, 1–19 (2005)

ReVelle, C., Hogan, K.: The maximum availability location problem. Transp. Sci. **23**, 192–200 (1989)

Toregas, C., Swain, R., ReVelle, C., Bergman, L.: The location of emergency service facilities. Oper. Res. **19**, 1363–1373 (1971)