A Hypercube Queuing Model Approach for the Location Optimization Problem of Emergency Vehicles for Large-Scale Study Areas

Felix Blank(✉)

University of Würzburg, 97070 Bavaria, Germany
Felix.Blank@uni-wuerzburg.de

Abstract. Emergency service systems provide essential services to people in need. Most of them have to operate under uncertainty and in complex environments. Their locations have to be chosen in a way that all or at least most of the incoming demand can be covered within a justifiable response time. In this paper, a hypercube queuing approach is presented that locates a high amount of emergency units within a large-scale study area. Due to the hypercube restrictions, computational times increase with the number of servers that are located. Therefore, an algorithm for the aggregation of demand areas is presented. To find an at least appropriate solution, a genetic algorithm is applied. It can be shown that computational times can be lowered significantly while the solution error is minimal. Furthermore, average response times for the emergency service system decrease with the location of additional servers in the study area.

Keywords: Hypercube · Emergency service system · Queuing

1 Introduction

Emergency service systems (ESS) are systems that provide immediate help to people in need. Possible areas of operation include fire-rescue, on-site medical care as well as emergency services in the case of man-made or natural disasters. Most ESS are subject to certain requirements regarding their coverage of the respective area and their response times. Due to the inherent uncertainties and the spatial distributions of demand in the system, the location decisions are not straightforward and have to be taken with respect to the underlying study area and the spatial distribution of the demand. Therefore, locations have to be chosen in a way that, even in remote parts of the study area, all or at least most of the incoming demand can be covered within a justifiable response time. A huge body of existing literature deals with the location optimization problem by using various models and techniques.

In this paper, a model is presented that incorporates methods of the hypercube queuing model (HQM) into a location problem for large-scale study areas and location decisions. The model can be used to optimize location decisions of ESS in large-scale study areas like the location of ambulances or other emergency vehicles in a city area. In order to deal with the computational efforts that are required for the location of many

M. Freitag et al. (Eds.): LDIC 2020, LNLO, pp. 321–330, 2020.
https://doi.org/10.1007/978-3-030-44783-0_31

emergency units, an aggregation algorithm (AA) is proposed. The paper is organized as follows: In Sect. 2, a brief literature review discusses existing contributions to the location of ESS and the HQM. Section 3 contains the formulation of the model as well as the AA. The applied metaheuristic, the exemplary study area as well as the computational results are presented in Sect. 4. Section 5 includes the conclusion as well as future research directions.

2 Literature Review

Facility location models are used by private and public organizations to determine optimal or near-optimal locations for their entities, like warehouses and, in the case of ESS, emergency stations or fire departments. Comprehensive reviews of facility location models for ESS can be found in Brotcorne et al. [1], Caunhye et al. [2] and more recently in Farahani et al. [3]. In general, facility location models for ESS can be divided into coverage models as well as p-median models. While the first group intends to locate facilities (also called servers in the case of ESS) in a way that maximizes the coverage over a demand area, the second group tries to minimize the distance between the demand points and the servers.

2.1 Coverage and P-Median Models

The very first contributions to the ESS location problem used static and deterministic inputs to obtain demand locations while ignoring factors like changing demands or other inherent dynamics of such systems. Toregas et al. [4] formulated the Location-Set-Covering-Model that required all demand being fulfilled within a pre-determined time frame. Church and ReVelle [5] proposed the Maximum-Coverage-Location-Problem (MCLP) that determines the location of each emergency unit in a way which maximises the covered space of each part of the study area. Daskin and Stern [6] stated an extension to the MCLP that maximizes the number of demand areas that is covered more than once. Hogan and ReVelle [7] formulated models that maximize backup coverage while Gendreau et al. [8] developed a model that uses two distinct time constraints. The p-median model was originally stated by Hakimi [9]. Calvo and Marks [10] used it to locate multi-level health facilities. Carson and Batta [11] determined a dynamic ambulance positioning strategy with the help of a p-median model.

2.2 Hypercube Queuing Model

The HQM is a markovian finite-state model and was initially stated by Larson [12] as a combination of queuing theory, facility location and analysis. It was then used to evaluate the performance of an underlying system. Based on a given set of server locations, the HQM can be used to derive certain performance measures that can be used to evaluate the decision-making. Larson [13] later introduced the approximate HQM that reduced computational difficulties while incorporating the original model. Since then the HQM has been widely applied and extended. This includes the better estimation of service rates [14], multiple dispatch of servers [15], modeling of co-located

servers [16] and customer-dependent service rates [17]. More recent extensions focus on the incorporation of waiting lines and customer preferences [18, 19].

Since the basic HQM and its extensions are descriptive models, they cannot be used to obtain optimal locations of facilities or servers [20]. The performance metrics that can be obtained by solving the basic HQM therefore need to be embedded in an optimization process [21, 22]. Batta et al. [23] and Saydam and Aytug [24] used the HQM in combination with the maximum expected coverage location problem (MEXCLP) to evaluate the performance of the derived locations. Galvão et al. [25] relax the server independence assumption of the maximum availability location problem with the help of the HQM. Geroliminis et al. [26] develop the spatial queuing model (SQM) and introduce server specific service rates as well as the allocation of demand areas to the responsible servers while minimizing the average response time of the overall system. Geroliminis et al. [27] develop a method for larger scale systems and propose a districting algorithm by reducing the steady states. Boyaci and Geroliminis [28] present two different models that also reduce the state-space by aggregating servers. Iannoni et al. [29] state a hypercube approximation algorithm to consider large numbers of emergency units. Akdogan et al. [30] propose different possible formulations of service rates in a SQM-based emergency service location study.

The use of the HQM can lead to more precise performance measures as well as the incorporation of server unavailability and backup structures. To the author's best knowledge, only [27, 28] and [29] consider large-scale ESS design while using HQM based methods. The existing body of literature reduces computational efforts mostly by reducing the state space. In large study areas the computational times do not only solely depend on the number of servers that are considered, but also on the number of demand areas. Since server responsibilities have to be checked for each state and demand area, increasing demand areas in a study area also increases computational times significantly. Due to the advances in computational power, larger number of emergency vehicles can nowadays be analysed without necessarily compromising the steady-state-space. Some of the required assumptions, like symmetrically located servers, identical workloads or homogenous demand, can reduce the accuracy of the solutions found by the model. In this paper, an approach that does not reduce the steady-state-space, but builds on dynamic formation of super demand areas is presented.

3 Large-Scale SQM

Consider a study area of J individual demand areas (atoms). In order to serve the incoming demand, several servers N have to be located within the study area. It is assumed that not every server can be sent to each atom. Therefore, each server has primary and lower level response areas that are determined with respect to the spatial distribution of the demand. The servers can only be busy or available and thus have only two, binary coded, states. If we consider a five-server system, in which the first, third and fifth server is busy, the corresponding state can be denoted as 10101. This generates 2^N different states for the system that are the vertices of the hypercube and are named B_a. The probabilities of each state are derived from an equation system that balances the flows between the separate hypercube states. The underlying equation

system is constructed by formulating one equation for each state that includes all upward and downward tranisitions. An upward transition happens for all incoming demand calls from the relevant demand areas and a downward transition happens whenever the server that differs between the two states completes its service. The resulting probabilities of the equation system describe the likelihood of each state of the hypercube queuing model.

The occurring demand is defined as a call for emergency help and happens solely at the center of each atom. It is assumed to be independent and not per definition identical between the atoms while following a time homogenous Poisson distribution. Whenever a demand (call) enters the system, the available servers are checked for availability and the closest available server is then dispatched. After completion of service, the dispatched server returns to the base location. If no server is available, the incoming call is lost to the system.

3.1 Model Formulation

The optimization model can then be formulated as follows:

$$\mathbf{min}\, T = \sum_{n=1}^{N} \sum_{j=1}^{J} p_{nj} t_{nj} \tag{1}$$

Subject to:

$$\sum_{j=1}^{J} f_j y_j \geq C_{cov} \tag{2}$$

$$\sum_{i \in W_j} x_i \geq y_j \forall j \in J \tag{3}$$

$$\sum_{i=1}^{I} x_i = N \tag{4}$$

$$x_i, y_j \in [0;1] \forall i \in I, j \in J \tag{5}$$

$$P\{B_b\}\left[\sum_{\left\{\begin{array}{c} a \\ B_a \in C_N : d_{ab}^- = 1 \end{array}\right\}} \lambda_{ab} + \sum_{\left\{\begin{array}{c} a \\ B_a \in C_N : d_{ab}^+ = 1 \end{array}\right\}} \mu_{ab}\right] = \sum_{\left\{\begin{array}{c} a \\ B_a \in C_N : d_{ab}^- = 1 \end{array}\right\}} P\{B_a\}\mu_{ab} + \sum_{\left\{\begin{array}{c} a \\ B_a \in C_N : d_{ab}^+ = 1 \end{array}\right\}} P\{B_a\}\lambda_{ab} \forall b = 0, 1, \ldots, 2^N - 1 \tag{6}$$

$$\sum_{a=0}^{2^N-1} P\{B_a\} = 1 \tag{7}$$

$$p_{nj} = f_j \frac{\sum_{B_a \in E_{nj}} P\{B_a\}}{1 - P\{2_{N-1}\}} \forall j \in J, n \in N \tag{8}$$

A notation similar to [30] is used. J describes the set of regions, N the number of servers to be located while I is the set of potential location sites. W_j is the set of

locations covering atom j. x_i, y_j are binary variables that show whether location site i is chosen or atom j is covered. C_{cov} is a pre-defined coverage value that takes a value equal or less to 1. λ is the system wide demand and f_j the demand fraction of atom j. p_{nj} is the fraction of dispatches server n sends to atom j, t_{nj} is the travel time of server n to atom j. C_N denotes the vertices of the N-dimensional hypercube. λ_{ab} and μ_{ab} are the upward and downward transition rates of the system, e.g. the transition rates that lead to a change of the system from state a to state b corresponding to the respective vertices B_a and B_b of the N-dimensional hypercube while $P\{B_a\}$ is the associated steady-state-probability of vertex a. λ_{ab} is the demand for a service offered by the ESS and μ_{ab} is the service rate for the requested service. d_{ab}^{+} and d_{ab}^{-} are the upward and downward Hamming distances from state a to state b and describe the difference in notation between the two states. For example, the difference (d_{ab}^{+}) between state 10000 and state 10001 is one. E_{nj} describes the set of states in which server n is the nearest available for region j. The model is controlled over the decision variables x_i and y_j with t_{nj} as an input.

Constraint (2) ensures that a certain pre-defined coverage level $(C_{cov} \leq 1)$ is met. Constraint (3) controls the decision variable y_j with respect to the coverage of Atom j. Constraint (4) guarantees that only the pre-defined number of vehicles is located. Constraint (6) specifies the equation system that is necessary to derive the steady-state-probability of the HQM. Each equation defines the balance of flows of one hypercube state. The sum of the probabilities of all hypercube states is equal to one (Constraint (7)). Since per definition incoming demand calls can be lost due to unavailable servers, the sum of the fraction of dispatches from all servers to one demand area can be lower than one. The denominator of (8) normalizes the fraction of dispatches under the consideration that not all servers are busy (the steady-state $P\{2_{N-1}\}$). Each fraction of dispatch p_{nj} describes the probability of a dispatch for server n to demand area j. This probability is then multiplied in (1) with the travel time of server n to demand area j to derive the expected average response time of the system.

The upward and downward transition rates are key inputs to the HQM and form the equations of the equation system in (6). Two important characteristics of the SQM as debuted by [26] are the spatial distribution of the demand as well as the assumption and calculation of districting levels. The later refers to the degree of coverage of each demand area that is provided by servers with downstream preferences. Since the model in this paper considers a large number of servers and demand areas, computational efforts for a complete backup are prohibitive and only third-level districting is used. The term d-th level districting refers to the partitioning of the study area in sub-areas according to the n-th nearest servers. For every $d > 1$ this means, that whenever the $d - 1$ nearest server is unavailable, the d nearest server responds. For every level of districting, the demand must be covered. The upward transition rate between vertex a and b of the hypercube can be calculated as follows:

$$\lambda_{ab} = \lambda_{kk}^{1} + \sum_{l_1 \in N: b_{l1}=1} \lambda_{l_1 k}^{2} + \sum_{m=2}^{M} \sum_{l_1 \ldots, l_m \in N: \prod_{i=1}^{m-1} b_{l_i}=1} \lambda_{l_1 k}^{m} \cap \lambda_{l_1 l_{m-1}}^{m-1} \cap \lambda_{l_1 l_{m-2}}^{m-2} \cap \ldots \cap \lambda_{l_1 l_2}^{2} \quad (11)$$

The equation above denotes that, given the system is in state a, server unit k responds to any demand in its area of responsibility D^{l}_{kk} or any other demand area D^{m}_{lk} for server l, as well as the $m - 1$ nearest responsible servers (denoted by l) are unavailable. λ^{l}_{kk} and λ^{m}_{lk} describe the demands of the demand area D^{l}_{kk} or D^{m}_{lk} respectively. The upward transition rate from state a to state b therefore consists of all demands from the demand areas for which server k is the primary server and from the demand areas for which server k is a lower tier server in the case of the first $m - 1$ servers of the preference list being unavailable. For a practical example of the partitioning of the study area into the sub areas the reader is referred to Akdogan et al. [30].

For calculating the downward transition rates, i.e. the service rate, there are several approaches in the literature. Geroliminis et al. [26] initially introduced a server-specific weighted-average approach that builds on the calculation of the demand rates without specifically considering the travel times to the atoms. In real life, travel times from the server to the demand area impact significantly the service rate of the server and the ESS. Akdogan et al. [30] therefore have stated an approach that is independent of the demand of the sub areas, but explicitly considers travel times from the server location to the location of the occurring demand. State a and state b differ at exactly one position of their state spaces. The location of the deployed server is then denoted by r_k with w_{kj} being the travel time from location r_k to demand area j. $\emptyset_{r_k}$ is the incident handling rate and T is the given time period. The incident handling rate per sub area then can be expressed as the number of possible deployments per hour with the denominator of (12) consisting of the incident handling time plus two times the mean travel times. The downward transition rate then is the sum of the incident handling rates of all sub areas $(j \in L_{jk})$ that can cause a transition from state a to state b:

$$\mu_{ab} = \sum\nolimits_{j \in L_{jk}} \frac{T}{\frac{T}{\emptyset_{r_k}} + 2 * w_{kj}} \tag{12}$$

3.2 Aggregation Algorithm

The HQM sets up a linear equation system with 2^N equations. The computational efforts and the solving time increase significantly with N. [27] tackle this problem with postulating a districting approach that reduces the number of equations to N. This is done by the assumption of symmetrical server locations as well as homogeneous demand and hence identical workloads of the servers. The states with the same number of busy servers are then summarized into one “super-state”. The approach presented in this paper does not compromise the expressiveness of the steady states but aggregates the demand areas dynamically with respect to the server locations as well as the demand areas. Since the exploratory study area considers a large number of demand areas, the computation time for generating the upward and downward transition rates increases significantly with the number of servers. Because the calculations have to be done for each proposed solution, the AA has to adapt dynamically to the server locations and the allocation of the demand areas to the servers. The AA is done in the following steps:

1. Determination of the D-nearest servers for each demand area (D = maximum level of districting)
2. Formation of new super demand areas for demand areas with the same server allocation for i = 1, …, D
3. Aggregation of demand of the original demand areas into demand of the super demand areas

Since the preferences of the individual demand areas do not change during the aggregation into super demand areas, the calculation of upward transition rates remains the same. The aggregation of the demand areas requires the calculation of new travel times. Equation (13) builds on (12) and uses a weighted demand approach to calculate the new travel times w_{kj}. The fraction of demand from demand region j of the super demand area A_s is used to weigh the original travel time from server k to demand area j. The downward transition rate is then calculated as a sum of all demand areas that belong to the respective super demand area.

$$\mu'_{ab} = \sum\nolimits_{j \in A_s \in L_{jk}} \frac{T}{\frac{T}{\emptyset_{r_k}} + 2 * \sum_{j \in A_s} \frac{\lambda_j}{\sum_{j \in A_s} \lambda_j} w_{kj}} \tag{13}$$

4 Solution Technique, Study Area and Results

The objective function in (1) has no closed-form expression. Therefore, an algorithm is needed to solve the proposed model. In this paper, a genetic algorithm (GA) as in [27] and [30] is applied. The GA tries to mimic evolutionary processes and to find optimal or near-optimal solutions by eliminating bad proposed solutions through survival of the fittest. In order to avoid local minima, mutation techniques are used.

4.1 Study Area

For the design of the exploratory study area, a 500 × 500 grid with 500 demand areas is considered. ESS often operate in urban environments, but also have responsibilities for more rural areas. Therefore, the demand areas are evenly spread over the study area and one urban agglomeration is considered. About a third of the demand points are located within the area of the urban agglomeration. Due to the combination of less and higher populated parts of the study area, the servers have to be located in a way that minimizes the mean response time for both groups of demand areas. The reference time period is one hour. The demand of the demand areas follows a time homogeneous Poisson distribution with the mean of 2 in the urban agglomeration area and 1 in the rural parts of the study area. $\emptyset_{r_k}$ is set to 1.

4.2 Results

The model, as well as the corresponding algorithm, were coded in C++ and run on a Intel Core i7 processor. The genetic algorithm was run 125 generations with a population size of 20 individuals in each.

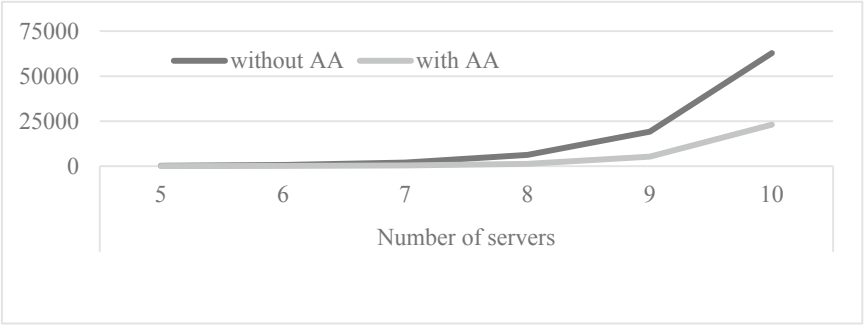

Fig. 1. Computation time in seconds.

The computation time increases significantly with the number of servers, especially in the nine and ten server case. The use of the proposed AA lowers the computation time to about one fourth in the 5 to 8 server-case and about one third in the 9 and 10 server-case. The number of super atoms rises with the servers that are considered due to the higher number of possible districting combinations that arise with more servers. This mitigates the performance advantage of the AA only to a rather small degree, as seen by the shallower course of the AA graph in Fig. 1.

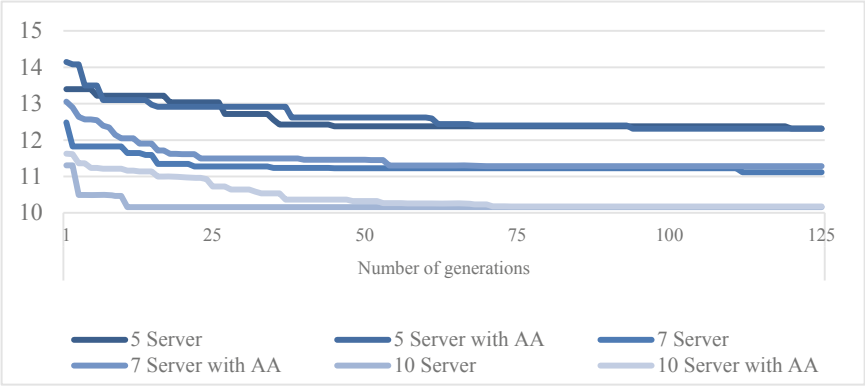

Fig. 2. Mean average response time in minutes for the 5, 7 and 10 server case.

The mean average response time of the system decreases significantly with the consideration of additional servers. With two (five) additional servers the mean response time of the system decreases by about 11(21)% due to the shorter travel times from the server locations to the demand areas. It can be shown that the marginal benefit of additional servers diminishes. The locations found in the optimization process with the AA are then used to compute the travel time when using the full model. The deviation percentage is not significant and under 2% in all cases (Figs. 2 and 3).

5 Server	6 Server	7 Server	8 Server	9 Server	10 Server
-0,12	1,85	1,47	0,45	0,1	0,15

Fig. 3. Percentage of deviation.

5 Conclusion

In this paper, an approach to incorporate larger study areas and larger number of servers into a hypercube queuing location model was presented. It could be shown that computation times can be reduced significantly while the steady state space as well as the quality of the solutions found is not compromised. Additional servers can also help to reduce the response time of ESS. It could be further shown that the quality of the solutions found is within 2% of the exact model. The proposed AA allows decision makers to include a larger number of servers or location sites into their analysis and decision process. Since the computational error is proven to be marginal, the use of proposed AA within HQM location models allows for a more accurate analysis and more realism, especially in the analysis of large study areas, like metropolitan areas.

Future research in this area could include the inclusion of dedicated waiting lines for incoming demand calls, the consideration of different day times as well as the comparison of different dispatch policies. The use of real data from ESS that have responsibilities for both rural and urban areas could add a further benefit.

References

1. Brotcorne, L., Laporte, G., Semet, F.: Ambulance location and relocation models. Eur. J. Oper. Res. **147**(3), 451–463 (2003)
2. Caunhye, A., Nie, X., Pokharel, S.: Optimization models in emergency logistics: a literature review. Socio-Econ. Plan. Sci. **46**(1), 4–13 (2012)
3. Farahani, R., Fallah, S., Ruiz, R., Hosseini, S., Asgari, N.: OR models in urban service facility location: a critical review of applications and future developments. Eur. J. Oper. Res. **276**(1), 1–27 (2019)
4. Toregas, C., Swain, R., ReVelle, C., Bergman, L.: The location of emergency service facilities. Oper. Res. **19**(1), 1363–1373 (1971)
5. Church, R., ReVelle, C.: The maximal covering location problem. Pap. Reg. Sci. Assoc. **32**(1), 101–118 (1974)
6. Daskin, M., Stern, E.: A hierarchical objective set covering model for emergency medical service vehicle deployment. Transp. Sci. **15**(2), 137–152 (1981)
7. Hogan, K., ReVelle, C.: Concepts and applications of backup coverage. Manage. Sci. **34**(11), 1434–1444 (1986)
8. Gendreau, M., Laporte, G., Semet, F.: Solving an ambulance location model by Tabu search. Locat. Sci. **5**(2), 75–88 (1997)
9. Hakimi, S.: Optimum locations of switching centers and the absolute centers and medians of a graph. Oper. Res. **12**(3), 450–459 (1964)
10. Calvo, A., Marks, H.: Location of health care facilities: an analytical approach. Socio-Econ. Plan. Sci. **7**(5), 407–422 (1973)

11. Carson, Y., Batta, R.: Locating an ambulance on the Amherst campus of the State University of New York at Buffalo. Interfaces **20**(5), 43–49 (1990)
12. Larson, R.: A hypercube queueing model for facility location and redistricting in urban facility service. Comput. Oper. Res. **1**(1), 67–95 (1974)
13. Larson, R.: Approximating the performance of urban emergency service systems. Oper. Res. **23**(5), 845–868 (1975)
14. Halpern, J.: The accuracy of estimates for the performance criteria in certain emergency service queueing systems. Transp. Sci. **11**(3), 223–241 (1977)
15. Chelst, K., Barlach, Z.: Multiple unit dispatches in emergency services: models to estimate system performance. Manage. Sci. **27**(12), 1390–1409 (1981)
16. Burwell, T., Jarvis, J., McKnew, M.: Modeling co-located servers and dispatch ties in the hypercube model. Comput. Oper. Res. **20**(2), 113–119 (1993)
17. Atkinson, J., Kovalenko, I., Kuznetsov, N., Mykhalevych, K.: A hypercube queueing loss model with customer-dependent service rates. Eur. J. Oper. Res. **191**(1), 223–239 (2008)
18. Souza, R., Morabito, R., Chiyoshi, F., Iannoni, A.: Incorporating priorities for waiting customers in the hypercube queuing model with application to an emergency service system in Brazil. Eur. J. Oper. Res. **242**(1), 274–285 (2008)
19. Rodrigues, L., Morabito, R., Chiyoshi, F., Iannoni, A., Saydam, C.: Towards hypercube queuing models for dispatch policies in queue and partial backup. Comput. Oper. Res. **84**, 92–105 (2008)
20. Galvão, R., Morabito, R.: Emergency service systems: the use of the hypercube queueing model in the solution of probabilistic location problems. Int. Trans. Oper. Res. **15**(5), 525–549 (2008)
21. Goldberg, J.: Operations research models for the deployment of emergency services vehicles. EMS Manage. J. **1**(1), 20–39 (2004)
22. Takeda, R., Widmer, J., Morabito, R.: Analysis of ambulance decentralization in urban emergency medical service using the hypercube queueing model. Comput. Oper. Res. **34**(3), 727–741 (2007)
23. Batta, R., Dolan, J., Krishnamurthy, N.: The maximal expected covering location problem: revisited. Transp. Sci. **23**(3), 277–287 (1989)
24. Saydam, C., Aytug, H.: Solving large-scale maximum expected covering location problems by genetic algorithms: a comparative study. Eur. J. Oper. Res. **141**(3), 480–495 (2002)
25. Galvão, R., Chiyoshi, F., Morabito, R.: Towards unified formulations and extensions of two classical probabilistic location problems. Comput. Oper. Res. **32**(1), 15–33 (2005)
26. Geroliminis, N., Karlaftis, M., Skabardonis, A.: A spatial queuing model for the emergency vehicle districting and location problem. Transp. Res. Part B **43**(7), 798–811 (2009)
27. Geroliminis, N., Kepaptsoglou, K., Karlaftis, M.: A hybrid hypercube – genetic algorithm approach for deploying many emergency response mobile units in an urban network. Eur. J. Oper. Res. **210**(2), 287–300 (2011)
28. Boyaci, B., Geroliminis, N.: Extended hypercube models for large scale spatial queuing systems. In: 91st Annual Meeting of the Transportation Research Board (2011)
29. Iannoni, A., Morabito, R., Saydam, C.: Optimizing large-scale emergency medical system operations on highways using the hypercube queuing model. Socio-Econ. Plan. Sci. **45**(3), 105–117 (2011)
30. Akdogan, M., Bayindir, Z., Iyigun, C.: Locating emergency vehicles with an approximate queuing model and a meta-heuristic solution approach. Transp. Res. Part C **90**, 134–155 (2018)