Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

A RECURSIVE OPTIMIZATION-SIMULATION APPROACH FOR THE AMBULANCE LOCATION AND DISPATCHING PROBLEM

Ettore Lanzarone
Enrico Galluccio

Istituto di Matematica Applicata e Tecnologie Informatiche (IMATI)
Consiglio Nazionale delle Ricerche (CNR)
Via Alfonso Corti 12
Milan, 20133, ITALY

Valérie Bélanger

Department of Logistics
and Operations Management
HEC Montréal
3000 Chemin de la Côte-Sainte-Catherine
Montréal, H3T 2A7, CANADA

Vittorio Nicoletta
Angel Ruiz

Department of Operations
and Decision Systems
Université Laval
2325 Rue de la Terrasse
Québec City, G1V 0A6, CANADA

ABSTRACT

The Ambulance Location and Dispatching Problem (ALDP) identifies the location of the available ambulances and the best dispatching policy to minimize the response times to answer the calls. However, the uncertain nature of the emergency calls makes it impossible to know in advance if the ambulance identified by the dispatching policy is available or not upon a call arrival. Thus, the probability that a vehicle is busy when a call arises, denoted as busy fraction, is usually considered in the literature. Probabilities can be estimated in several manners, but simulation seems to be well suited for this purpose. In this work, we propose four Recursive Optimization-Simulation Approaches to estimate the ALDP busy fraction, and we apply them to a set of realistic instances. Numerical results confirm that the most sophisticated and computing demanding approaches offer a better performance.

1 INTRODUCTION

Emergency Medical Services (EMSs) are crucial in health care systems. They provide out-of-hospital acute medical care and transportation to the appropriate health center for injured and ill people. Organizing and managing the service is an extremely challenging task, as EMSs manage a large variety of resources (e.g., health care personnel, ambulances and call centers) and face difficult challenges related to the uncertainty of the emergency calls. Indeed, calls randomly arrive from different areas of the served territory and must be served quickly: every second is important in life threatening situations. In this context, EMS organizations must answer two important questions, i.e., *where should the ambulances be located?* and *which ambulance should take care of the arriving call?*

These decisions are the aim of the Ambulance Location and Dispatching Problem (ALDP), which identifies the location of the available ambulances and provides a dispatching policy to minimize the

expected response time to answer the calls (Bélanger et al. 2015). In particular, the ALDP defines an ordered *dispatching list* of ambulances for each demand zone, so that, when a call arises from a zone, the first available ambulance in its dispatching list is sent to answer the call. If no vehicles in the list are available, the closest available ambulance is sent. Finally, if no vehicle is available at all, the call is redirected to another emergency service. Therefore, the ALDP merges tactical and operational decisions. Tactical decisions can be considered as static, as they do not change during the planning horizon. In fact, once the locations of the ambulances are decided, vehicles are assigned to them and come back every time as soon as they complete their mission. On the other hand, dispatching decisions are rather dynamic, since they require the knowledge of the ambulances that will be available when a call arrives. However, call arrivals are random, and it is almost impossible to know the busy ambulances in advance.

An effective approach to handle this uncertainty is to model ambulances' availabilities in a probabilistic manner, but doing so is far from being trivial and requires a fair estimation of ambulances' probabilities to be busy. In the literature, a well-recognized approach is to describe such probabilities in function of the so-called *busy fraction*, i.e., the ratio of time in which an ambulance is serving a call and not available to serve another call (Daskin 1983). However, also in this case, a fair estimation of the busy fractions is needed. The EMS literature has relied for a long time on queueing theory to estimate busy fractions, and a large number of models are available. However, most of these analytical models tend to be very difficult and often require assumptions that are hardly acceptable in real-life contexts.

Simulation, on the other hand, seems to be well suited to estimate busy fractions. Anyway, although simulation has been largely used in the EMS literature, in most of the cases its purpose was only to evaluate the performance of a plan, usually produced by an optimization model (Fujiwara et al. 1987; Harewood 2002). Just recently, a new trend is gaining more popularity in EMS, especially for location problems: optimization for simulation (Aboueljinane et al. 2013). This approach consists of selecting some deployment strategies to be tested in the simulator (Silva and Pinto 2010).

In this work, we propose a Recursive Optimization-Simulation Approach (ROSA) to model and estimate ambulance busy fractions. The ROSA exploits simulation not only to validate a solution, but also to recursively improve the optimization model. We propose four variants of the ROSA and we analyze their performance. To the best of our knowledge, ROSA has never been applied to ALDP nor to more general EMS optimization problems.

The rest of the paper is structured as follows. Section 2 briefly presents the two ingredients of the approach, i.e., the ALDP and the simulator. Section 3 proposes the first three variants of the ROSA, while Section 4 reports their empirical results. In addition, Section 5 presents a fourth variant that further improves the outcomes with respect to the others. Finally, Section 6 draws some conclusions.

2 OPTIMIZATION AND SIMULATION MODELS

This work considers a version of the ALDP in which *i*) relocation is neglected, and *ii*) each standby site may host up to one ambulance, which is a reasonable assumption when the number of standby sites is much larger than the number of ambulances, as it is in most of real cases, including our numerical tests.

In the following, we briefly present the ALDP formulation, which is used to locate ambulances and define the dispatch policy, and then we describe the simulator, which is used to execute a solution.

2.1 ALDP formulation

The ALDP is defined on a graph $G = (V,E)$ with $V = I \cup J$, $I = \{v_1, \ldots, v_n\}$, $J = \{v_{n+1}, \ldots, v_{n+m}\}$, and $E = \{(v_i, v_j) : v_i, v_j \in V\}$. I is the set of demand zones, J the set of potential standby sites, and E the set of edges, where each edge has associated a traveling time t_{ji}. A demand zone v_i is characterized by a centroid and an overall demand d_i over the planning horizon. A potential standby site is a site within a zone where one ambulance can be located while waiting for calls. The set of available vehicles is denoted by K, and the set of positions in the dispatching list of each zone by Z. For each ambulance, we define

the maximum workload W as the maximum number of missions a vehicle can answer during the planning horizon, and we denote by q the busy fraction; both of them are the same for all vehicles.

Two sets of decision variables are used. Variable x_j is equal to 1 if a vehicle is located in $v_j \in J$ (0 otherwise), and variable y_{ij}^z is equal to 1 if a vehicle located in $v_j \in J$ is in position z of the dispatching list of $v_i \in I$ (0 otherwise). The goal is to minimize the sum of the response times to answer to the calls.

The ALDP is formulated as:

$$\min \sum_{i\in I}\sum_{z\in Z}\sum_{j\in J}(1-q)\,q^{z-1}d_i t_{ji} y_{ij}^z \tag{1}$$

subject to

$$\sum_{j\in J} x_j = |K| \tag{2}$$

$$\sum_{z\in Z}\sum_{i\in I}(1-q)\,q^{z-1}d_i y_{ij}^z \le W \qquad \forall j \in J \tag{3}$$

$$\sum_{z\in Z} y_{ij}^z \le 1 \qquad \forall j \in J, i \in I \tag{4}$$

$$\sum_{j\in J} y_{ij}^z \le 1 \qquad \forall z \in Z, i \in I \tag{5}$$

$$y_{ij}^z \le x_j \qquad \forall z \in Z, i \in I, j \in J \tag{6}$$

$$x_j \in \{0,1\} \qquad \forall j \in J \tag{7}$$

$$y_{ij}^z \in \{0,1\} \qquad \forall z \in Z, i \in I, j \in J \tag{8}$$

Constraints (2) impose that the $|K|$ vehicles are located in the standby sites. Constraint (3) ensure that the demand assigned to a vehicle is less than or equal to W. Finally, a vehicle cannot occupy more than one position in the dispatching list of a given demand zone – constraints (4) – and exactly one vehicle appears at each position of the dispatching list of a given demand zone – constraints (5). Additional details can be found in (Bélanger et al. 2015).

Let us now focus on the term $(1-q)q^{z-1}$ in (1) and (3), which accounts for the probability that the ambulance in position z of a dispatching list of a zone answers the call. In fact, the first ambulance in the dispatching list answers the call only if it is idle, i.e., with probability $1-q$. The second ambulance of the list answers the same call only if it is idle and if the first ambulance is busy, i.e., with probability $(1-q)q$. The third ambulance of the list answers the call with probability $(1-q)q^2$, and so on.

With the locations of the ambulances and the dispatching lists, it is possible to compute the Expected Response Time (ERT), which also includes the potential contribution of the ambulances that are not in the dispatching list, as well as a penalty T_p if no ambulance is available. To this end, we define an *extended list* for each zone, whose length is $|K|$: the first $|Z|$ elements are given by the dispatching list and, starting from position $|Z|+1$, the remaining ambulances are ordered from the closest to the farthest. Extended lists are defined by variables $\widetilde{y}_{ij}^z$, where $\widetilde{y}_{ij}^z = y_{ij}^z$ for $z \le |Z|$, and for $z > |Z|$ we state that $\widetilde{y}_{ij}^z = 1$ if and only if an ambulance is located in v_j and is the closest remaining ambulance to zone v_i. Then:

$$ERT = \sum_{i\in I} ERT_i = \sum_{i\in I}\sum_{z\in Z}\sum_{j\in J}(1-q)\,q^{z-1}d_i t_{ji}\widetilde{y}_{ij}^z + \sum_{i\in I} d_i q^{|K|} T_p \tag{9}$$

2.2 Simulator

We consider a Discrete Event Simulation (DES) model, which contains two types of entities (i.e., calls and vehicles) and two types of events.

A vehicle is either *idle* or *busy*, and it can be assigned to a call only if idle. Whenever a vehicle is assigned to a call, its status switches to busy for a service time given by the sum of the response time (the

travel time from its standby site to the call zone) and the working time (which includes transportation to the hospital and return to the standby site). Although the working time is quite regular, it slightly varies from mission to mission. However, since this variability is limited and would impact all the proposed methods in a similar manner, we assume a constant working time to better focus on the comparison of the proposed ROSA variants. In particular, we set all working times to 4320 seconds (i.e., 1.2 hours).

The first type of event (dispatch) occurs when a call arises in a demand zone and is assigned to an idle ambulance; the second type of event (release) happens when a vehicle comes back to its standby site after having completed a service and becomes idle. In fact, we consider that each vehicle returns back to its standby site, defined by the ALDP, without being relocated or rerouted to serve another call. Dispatch events follow the dispatching policy produced as a solution by the ALDP.

The simulator uses as an input a *demand scenario*, which is in terms of a *call list* generated as follows. We consider the demand d_i in each zone i, and we convert this information into rates per second to obtain the vector $\Lambda = (\lambda_1, \lambda_2, \ldots, \lambda_{|I|})$, where λ_i is the average number of calls per second in zone i. Then, we use a Poisson model to exploit the memoryless property of the interarrival times, i.e., the time interval in seconds until the next call is modeled via an exponential variable with parameter $\lambda_{TOT} = \sum_{i=1}^{N} \lambda_i$. Finally, to assign a zone to each call, we consider the following probabilities:

$$\Pr\{\text{call arises from zone } i\} = \frac{\lambda_i}{\lambda_{TOT}}.$$

Several scenarios, associated with as many call lists, are considered to compute average values and confidence intervals for the system variables. The simulator also uses as inputs the ALDP solution (i.e., the location of each vehicle k and the dispatching list for each zone i) and the travel times t_{ji}, which are assumed known and deterministic. The main outputs of the simulator are the busy fraction, the response time and the working time of each vehicle k. The pseudocode of the simulator is presented in Algorithm 1.

Algorithm 1 Simulator

```
for each simulated scenario do
    Initialize response time and working time to 0 for each vehicle
    Initialize each vehicle as idle
    while event set is not empty do
        Look for the next event
        if next event is a call from zone i then
            Send vehicle k (the first available for zone i according to the extended list)*
            Switch vehicle k status from idle to busy
            Create the future event: "vehicle k finishes the service"
        else if next event is a finish service for vehicle k then
            Switch vehicle k status from busy to idle
            Add the response time and the working time to ambulance k
    Calculate the busy fraction of each ambulance for the scenario
    Calculate global response time and global working time for the scenario
    Calculate other relevant data
Extract mean and confidence intervals of busy fractions
Extract mean and confidence intervals of global response time and global working time
Extract other relevant statistics
```

* *if there are no idle vehicles when a call is received, the model deals with it by adding a penalty*

3 ROSA-BASED APPROACHES

We present here the first three variants of the ROSA, and discuss their advantages and disadvantages in terms of complexity and accuracy.

3.1 BRM

In the basic ROSA model (BRM), we consider the ALDP formulation presented in Section 2.1, and we only tune the value of the busy fraction q, which is the same for all ambulances. We remark that the term $(1-q)^{z-1}$ relies on assuming independence among the ambulances. Thus, the two assumptions of independence and equal q values are at the basis of the BRM.

We run the ALDP with an initial q_0; then, we run the simulator to compute an empirical q as the average over the simulated scenarios. This new q is then used for solving the ALDP again. We iterate the process up to convergence, i.e., until the busy fraction does not change between two consecutive iterations. In practice, we consider that convergence is reached when the difference between two busy fractions is lower than a tolerance $\varepsilon = 10^{-5}$.

3.2 PSSM and QTSSM

To relax the independence assumption of the BRM, we generalize the ALDP model of Section 2.1 as follows:

$$\min \sum_{i\in I}\sum_{z\in Z}\sum_{j\in J} \xi_z d_i t_{ji} y_{ij}^z \tag{10}$$

subject to

$$\sum_{z\in Z}\sum_{i\in I} \xi_z q^{z-1} d_i y_{ij}^z \le W \qquad \forall j \in J \tag{11}$$

$$Constraints\ (2),\ (4)-(8)$$

In particular, the term $(1-q)q^{z-1}$ in (1) and (3) is replaced by a parameter ξ_z for each $z \in Z$. Accordingly, also (9) is modified as follows:

$$ERT = \sum_{i\in I}\sum_{z\in Z}\sum_{j\in J} \xi_z d_i t_{ji} \widetilde{y}_{ij}^z + \sum_{i\in I} d_i \left(1 - \sum_{z\in Z} \xi_z\right) T_p \tag{12}$$

Now, the optimization model and the simulator exchange the set of parameters ξ_z instead of the single parameter q. For the rest, the iterative optimization-simulation scheme is the same. As for the initialization, we start with $\xi_z^0 = (1-q_0)q_0^{z-1}$. The convergence of the optimization-simulation framework cannot rely on a single parameter as in the BRM. Here, convergence is reached when $L_k^{[\iota]} = L_k^{[\iota+1]}$ and $D_j^{[\iota]} = D_j^{[\iota+1]}$ $\forall k, j$, where $L_k^{[\iota]}$ denotes the locations generated by the ALDP at iteration ι and $D_j^{[\iota]}$ the corresponding dispatching lists.

We consider two different methods to estimate the parameters ξ_k. The former relies only on simulation and combinatorics (Section 3.2.1); the latter is based on both queuing theory and simulation (Section 3.2.2). In both cases, the assumption of equal q among the ambulances still persists.

3.2.1 PSSM

The first estimation method is referred to as Pure Simulation Sampling Model (PSSM).

All vehicles have the same probability q of being busy; thus, given two groups G1 and G2 with n ambulances, the probability that all ambulances in G1 are busy is the same as the probability that all ambulances in G2 are busy. We refer to this probability as ψ_n.

If we randomly select a time instant and observe that the number of busy ambulances is b, then the probability that an ambulance k is busy is given by:

$$\Pr\{\text{ambulance } k \text{ is busy} \mid b \text{ ambulances are busy}\} = \frac{b}{|K|}. \tag{13}$$

Similarly, the probability that a given n-uple of ambulances are all busy is:

$$\Pr\{\text{a given } n\text{-uple of ambulances are all busy} \mid b \text{ ambulances are busy}\} = \frac{\binom{b}{n}}{\binom{|K|}{n}} \tag{14}$$

If we know that b ambulances are busy, ψ_n is given by (14). Finally, we derive ξ_z based on ψ_z:

$$\begin{aligned} \xi_z &= \Pr\{\text{ambulance } z+1 \text{ is idle and ambulances 1 to } z \text{ are busy}\} = \\ &= \Pr\{\text{first } z \text{ ambulances are busy}\} - \Pr\{\text{first } z+1 \text{ ambulances are busy}\} = \psi_z - \psi_{z+1} \end{aligned} \tag{15}$$

We can thus estimate ξ_z from the simulator. To this end, we randomly sample 400 time instants for each simulation scenario. For each time instant and scenario, we get the number b of busy ambulances, we compute with b the value of ψ_z for each $z \in Z$, and we finally compute the corresponding ξ_z. Then, for each ψ_z, we get the average value over the time instants and the scenarios.

3.2.2 QTSSM

The second estimation method is referred to as Queuing Theory and Simulation Sampling Model (QTSSM), as it uses both simulation and queuing theory to improve the estimation.

We refer to two literature works. (Larson 1974) solved the location problem for EMS using an exact hypercube model, in which each vertex of the hypercube is associated with a particular combination of busy vehicles. The Markov properties of the Poisson interarrival times were then exploited to generate the so-called balance equations. While this method provides an exact solution, generating and solving the balance equations is extremely demanding from a computational viewpoint and becomes infeasible as the number of vehicles increases. Thus, (Larson 1975) used the conditional probability to derive a correction factor for the non-independence of the vehicles. In particular, we use the following expression that (Budge et al. 2009) derived from (Larson 1975):

$$\xi_z \simeq Q(|K|,\rho,z)\,(1-q)q^{z-1} \tag{16}$$

$$Q(|K|,\rho,z) = \frac{\pi_0}{|K|!\left(1-\rho\left(1-\pi_{|K|}\right)\right)} \frac{(|K|-z)!}{\left(1-\pi_{|K|}\right)^{z-1}} \sum_{u=z-1}^{|K|-1} \frac{(|K|-u)|K|^u \rho^{u-z+1}}{(u-z+1)!} \tag{17}$$

where π_n are the steady-state probabilities and ρ is the occupation rate per server. It is worth noting that Larson estimated all parameters using conditional probability and queuing theory results, by solving a set of non-linear simultaneous equations, without relying on simulation.

To estimate the parameters, we look at our problem from a queuing theory perspective and we model it as an Erlang loss system (Gross and Harris 2008), where the emergency calls are the customers who enter the system and the ambulances are the servers. Servers are busy when they are dealing with a call (during response and working times). We have already assumed that the call is lost and a penalty is incurred when all ambulances are busy, and we have already stated that the calls follow a Poisson process. The process is thus modeled as an M/G/$|K|$/$|K|$ queue, where M indicates that the interarrival times are exponential, and G that the service times follow any distribution. $|K|$ is the number of servers, equal to the maximum number of customers handled by the system (in our case, the maximum number of calls that can be served in a given time instant).

Table 1: Summary of the tested alternatives (times in seconds).

	$\|I\|$	$\|J\|$	$\|K\|$	$\|Z\|$	T_p	W	d_i		t_{ji}	
							min	max	min	max
SI–2	30	30	2	2	420	5000	41	496	15	557
SI–3	30	30	3	2	420	5000	41	496	15	557
SI–4	30	30	4	2	420	5000	41	496	15	557
MI–12	149	48	12	2	420	5000	16	737	15	1174
LI–60	595	218	60	2	420	5000	1	1715	15	3124

We define the occupation rate $a = \frac{\lambda}{\mu}$ and compute the occupation rate per server ρ as $\frac{a}{|K|} = \frac{\lambda}{|K|\mu}$. ρ is very similar to the busy fraction q, but it also takes into consideration the calls that are lost ($q \leq \rho$). Then, the steady state probability π_n of state n is the probability that n clients will be in the system and that n servers will be busy in the long run. Based on the Markov chains theory (Jarvis 1985), we have:

$$\pi_n = \frac{\frac{a^n}{n!}}{\sum_{j=0}^{|K|} \frac{a^j}{j!}} \tag{18}$$

Notice that, although this formula is derived for the M/M/$|K|$/$|K|$ queue, the insensitivity or robustness property of queueing systems (Burman 1981) states that the distribution form of the service times is irrelevant; thus, it can also be applied to the M/G/$|K|$/$|K|$ queue, provided that the mean service time μ is known. Moreover, let us observe that in our case the service times include both the response and the working times. While the latter is assumed to be constant, the response time is variable and depends on the ALDP solution. Anyway, its relative contribution with respect to the working time is negligible, and we may consider a fixed mean service time μ. We can now can calculate $Q(|K|, \rho, z)$ based on λ and μ. As for λ, it is the known parameter λ_{TOT} of the Poisson model of the interarrival times. μ is estimated empirically as the average of the sum of response and working times across all trips over all simulations. Finally, we also compute q from simulation, as in BRM, and we obtain $\xi_z \ \forall z \in Z$.

4 RESULTS

The main goal of this section is to assess the accuracy of the ROSA variants proposed in Section 3. To this end, we compare the *ERT* to the corresponding empirical response time from simulation. More precisely, we consider the empirical total transportation and penalty time in each scenario and we compute its average value over the scenarios, which is denoted as Simulated Response Time (*SRT*). Moreover, we consider the probability θ_o^z that the ambulance in position z of an extended list answers a call; with that, we derive the contribution θ_o^l of the $|Z|$ ambulances in the dispatching list, the contribution θ_o^{nl} of the other ambulances, and the penalty contribution θ_o^p. As for the optimization, they are derived as follows:

$$\theta_o^z = \xi_z \qquad \theta_o^l = \sum_{z \in Z} \theta_o^z \qquad \theta_o^p = 1 - \sum_{z \in Z} \xi_z \qquad \theta_o^{nl} = 1 - \theta_o^l - \theta_o^p$$

while, for the simulation, the corresponding probabilities θ_s^z, θ_s^l, θ_s^{nl} and θ_s^p are simply sampled.

Tests are run considering pseudo-real data from the city of Montréal and the near suburb of Laval, Québec, Canada. The territory is divided into 595 demand zones, and both demands d_i and travel times t_{ji} are those in (Bélanger et al. 2015). Additional information about the data can be found in (Kergosien et al. 2015). The entire territory is considered for the Large Instance (LI). Moreover, we define a Small Instance (SI) and a Medium Instance (MI), which include 30 and 149 zones in downtown Montréal, respectively. Three cardinalities of the ambulance set are considered for SI, while only one for MI and LI, thus giving 5 alternatives. Their characteristics are summarized in Table 1. The average numbers λ_i of calls per second in all zones i, used to generate the call lists for the simulator, are taken according to d_i. With them, we randomly generate 500 random scenarios.

Table 2: Results of BRM. Values of q, *ERT* and *SRT* at the initial iteration and at convergence, together with difference $SRT - ERT$ as percentage of *ERT* (a). θ_o^1 vs θ_s^1; θ_o^2 vs θ_s^2; θ_o^{nl} vs θ_s^{nl}; θ_o^p vs θ_s^p (b). Values of *ERT* and *SRT* are expressed in seconds.

	q		*ERT*		*SRT*		Difference	
	initial	converg	initial	converg	initial	converg	initial	converg
SI–2	0.5	0.43	1555747	1430990	1515390	1515390	–2.66%	+5.90%
SI–3	0.5	0.34	1284635	1039699	1112726	1112726	–13.38%	+7.02%
SI–4	0.5	0.27	1107145	831067	881520	872687	–20.38%	+5.01%
MI–12	0.5	0.57	9030956	9827561	10867120	10867120	+20.33%	+10.58%
LI–60	0.5	0.53	42390570	44035665	50529180	50508731	+19.20%	+14.70%

(a)

	θ_o^1	θ_s^1	θ_o^2	θ_s^2	θ_o^{nl}	θ_s^{nl}	θ_o^p	θ_s^p
SI–2	0.566	0.561	0.246	0.209	0.000	0.000	0.189	0.230
SI–3	0.655	0.653	0.226	0.195	0.078	0.073	0.041	0.079
SI–4	0.726	0.724	0.199	0.176	0.069	0.079	0.006	0.021
MI–12	0.434	0.425	0.246	0.202	0.320	0.347	0.001	0.026
LI–60	0.470	0.409	0.249	0.192	0.281	0.399	0.000	0.000

(b)

The ALDP model is solved with CPLEX 12.7 on a Microsoft Windows machine equipped with processor Intel Core i7-4700MQ @ 2.40 GHz and 8 GB of installed RAM. The computational time of each run is always below 1 second for SI and MI, while below 1 minute for LI.

We always set $q_0 = 0.5$ as in (Bélanger et al. 2015).

Convergence is reached very quickly in all cases, i.e., after 2 to 6 iterations. Since convergence is based on a quantitative comparison between optimization and simulation results, a cyclic behavior occurs in PSSM and QTSSM. However, in all the cases we solved, the cyclic oscillations of *ERT* and *SRT* values are minimal and their impact on the solution negligible. Thus, any solution of the cyclic behavior can be the outcome of the optimization-simulation framework.

Detailed results are reported below for each ROSA variant.

4.1 BRM

Results in Table 2a show that the difference between the ERT (produced by the optimization) and the SRT (produced by the simulation) is improved at convergence in all cases but one. We remind that the initial solution with an *a priori* defined $q = 0.5$ corresponds to the approach in (Bélanger et al. 2015). However, despite of this improvement, the ERT to SRT gap is still too large, ranging from 5.01% up to 14.70%.

To investigate the reasons of this gap, we compare θ_o^1, θ_o^2, θ_o^{nl}, θ_o^p with θ_s^1, θ_s^2, θ_s^{nl}, θ_s^p in Table 2b. We observe a general overestimation of the probability that the second ambulance in the dispatching list answers a call. Thus, we may conclude that, while the term $1-q$ is a good estimate of the probability that an ambulance is idle, $(1-q)q^{z-1}$ for $z > 1$ is not a good estimate. Moreover, we observe that, for LI, also the probability that the first ambulance in the dispatching list answers a call is overestimated.

4.2 PSSM and QTSSM

The results of PSSM are reported in Table 3a. It can be observed that the PSSM provides a significant improvement in ERT accuracy for all instances. In fact, the ERT to SRT gap is lower than 2.44% for all small instances, and equal to 5.79% for MI. Nevertheless, the gap is still high for LI and, in this case, also the PSSM overestimates the probability that the ambulances in the dispatching lists answer a call (Table 3b). The results of QTSSM are reported in Table 4; they are very similar to those obtained with PSSM.

Table 3: Results of PSSM. Values of q, ERT and SRT at convergence, together with difference $SRT - ERT$ as percentage of ERT (a). θ_o^1 vs θ_s^1; θ_o^2 vs θ_s^2; θ_o^{nl} vs θ_s^{nl}; θ_o^p vs θ_s^p (b). Values of ERT and SRT are expressed in seconds.

	q	ERT	SRT	Difference
SI–2	0.43	1479247	1515390	+2.44%
SI–3	0.34	1091123	1112925	+2.00%
SI–4	0.27	860079	872687	+1.47%
MI–12	0.57	10271879	10867120	+5.79%
LI–60	0.57	44472172	50509121	+13.57%

(a)

	θ_o^1	θ_s^1	θ_o^2	θ_s^2	θ_o^{nl}	θ_s^{nl}	θ_o^p	θ_s^p
SI–2	0.565	0.561	0.205	0.209	0.000	0.000	0.229	0.230
SI–3	0.656	0.654	0.196	0.194	0.070	0.073	0.078	0.079
SI–4	0.726	0.724	0.180	0.176	0.073	0.079	0.021	0.021
MI–12	0.434	0.425	0.224	0.202	0.316	0.347	0.026	0.026
LI–60	0.470	0.409	0.244	0.192	0.285	0.399	0.000	0.000

(b)

4.3 Comparison

We may first observe that all of the three approaches lead to the same exact SRT in SI with 2 vehicles and in MI. On the contrary, the solutions are slightly different in the other cases. In SI with 3 vehicles, BRM and QTSSM share the same solution, while PSSM has a higher (worse) SRT value. In SI with 4 vehicles, BRM and PSSM share the same solution, while QTSSM has a lower (better) SRT value. Finally, all solutions are different in LI. These differences show that the ALDP makes different decisions depending on the q value, confirming that overestimating or underestimating q may lead to different system performance, and that the ERT values may sometimes seem very attractive but unreliable. To confirm this last statement, we observe that the response time is always underestimated in the optimization, especially for BRM (positive differences in Tables 2a, 3a, 4a). To conclude, QTSSM solutions are better for two reasons: they have lower SRT values and the planned ERT is closer to the simulated SRT.

Further analyses are performed to compare the confidence intervals of the parameters extracted from simulation. Tables 3b and 4b show very similar values (the difference is less than 1% for each parameter). However, these values are averaged over 500 scenarios. If we consider the variance of the probabilities over the scenarios, we may observe much higher values for PSSM.

We recall here that, in PSSM, the probabilities are estimated by sampling 400 time instants; thus, the variance could be reduced by increasing the number of samples, but empirical studies have shown that this only increases the computational time while the performance would not be better than for QTSSM.

The lower variance of QTSSM has the following relevant advantage: QTSSM is more stable than PSSM, thus avoiding cases with impaired quality of the solution. To prove that, we have run the two models several times using different seeds for the simulation. We have observed that the results obtained by QTSSM are basically the same for each seed, while for PSSM the probabilities slightly change every time, leading to small differences in the ERT and sometimes causing different ALDP solutions.

In conclusion, QTSSM almost always provides a better solution for the ALDP, both in terms of efficiency and precision, and it is more stable than PSSM. However, this is not true for LI, where the solution provided by QTSSM at convergence is worse than the BRM one in terms of SRT. Moreover, the gap between ERT and SRT is still large. This is why, in Section 5, we propose a fourth approach that aims to overcome the drawbacks showed by QTSSM in LI, mainly due to the deviation of θ_o^1 with respect to θ_s^1.

Table 4: Results of QTSSM. Values of q, ERT and SRT at convergence, together with difference $SRT - ERT$ as percentage of ERT (a). θ_o^1 vs θ_s^1; θ_o^2 vs θ_s^2; θ_o^{nl} vs θ_s^{nl}; θ_o^p vs θ_s^p (b). Values of ERT and SRT are expressed in seconds.

	q	ERT	SRT	Difference
SI–2	0.43	1480670	1515390	+2.34%
SI–3	0.34	1093288	1112726	+1.78%
SI–4	0.27	860188	872638	+1.45%
MI–12	0.57	10283386	10867120	+5.68%
LI–60	0.57	44482173	50508791	+13.55%

(a)

	θ_o^1	θ_s^1	θ_o^2	θ_s^2	θ_o^{nl}	θ_s^{nl}	θ_o^p	θ_s^p
SI–2	0.566	0.561	0.204	0.209	0.000	0.000	0.230	0.229
SI–3	0.656	0.653	0.196	0.195	0.070	0.073	0.079	0.079
SI–4	0.726	0.724	0.180	0.175	0.073	0.079	0.021	0.021
MI–12	0.434	0.425	0.224	0.202	0.316	0.347	0.027	0.026
LI–60	0.470	0.409	0.244	0.192	0.285	0.399	0.000	0.000

(b)

5 E-QTSSM

We recall that the main assumption of QTSSM is to consider the same busy fraction q for all vehicles. To evaluate the pertinence of this assumption, we analyze the simulated busy fraction q_k of each ambulance k in QTSSM at convergence. Table 5 shows that the q_k values are homogeneous in all small instances, with a difference $\max_k\{q_k\} - \min_k\{q_k\}$ of at most 0.05. However, such difference increases with the size, being 0.17 in MI and 0.65 in LI. This q_k heterogeneity can be the reason for the observed gap between θ_o^1 and θ_s^1. Thus, to improve the outcomes, we need to tackle this heterogeneity and, in particular, we should consider the fact that busier ambulances usually serve high-demand areas.

Our first attempt is to replace the average q with a weighted average $\overline{q}$, where the weights are given by the number of calls received by the zones of competence of each ambulance k. In this way, busier ambulances from zones with higher demands contribute more to $\overline{q}$. However, these weights are hard to compute from the data produced by the simulator. Therefore, as busier ambulances with higher q_k serve high-demand areas and ambulances with lower q_k serve more peripheral areas, we adopt a simple yet effective approach that assigns weights proportionally to q_k. Thus, $\overline{q} = \sum_{k\in K} \alpha_k q_k$ with $\alpha_k = \frac{q_k}{\sum_{k\in K} q_k}$.

Afterwards, we work as in QTSSM, but with $\overline{q}$ instead of q, i.e., $\xi_z \simeq Q(|K|,\rho,z)(1-\overline{q})\overline{q}^{z-1}$.

We refer to this new approach as Enhanced QTSSM (E-QTSSM). The same criterion for convergence of PSSM and QTSSM is adopted.

5.1 Results

We report in Table 6 the results of E-QTSSM only for LI, for which QTSSM shows limitations. We observe that E-QTSSM reduces the ERT to SRT gap, passing from the 13.55% of QTSSM to the 3.80% (Table 6a), and the deviation of θ_o^1 with respect to θ_s^1 (Table 6b). Convergence is reached very quickly (after 2 to 6 iterations) also in this case. In the other instances, outcomes are similar between QTSSM and E-QTSSM.

6 CONCLUSIONS

In this work, we implement several optimization-simulation frameworks to solve the ALDP, with the aim of improving the busy fraction approximation. The goal is twofold: on the one hand, we make the system more effective by lowering the SRT; on the other hand, we ensure that the model is a correct approximation of reality by reducing the gap between SRT and ERT.

Table 5: Simulated q_k values of all ambulances $k \in K$ at convergence for QTSSM.

	List of q_k values
SI–2	0.41; 0.46
SI–3	0.34; 0.34; 0.36
SI–4	0.26; 0.26; 0.28; 0.30
MI–12	0.45; 0.50; 0.50; 0.56; 0.53; 0.57; 0.58; 0.58; 0.61; 0.62; 0.63; 0.65
LI–60	0.12; 0.19; 0.24; 0.24; 0.25; 0.26; 0.33; 0.34; 0.35; 0.36; 0.36; 0.36; 0.38; 0.39; 0.39; 0.39; 0.41; 0.42; 0.42; 0.43; 0.46; 0.46; 0.47; 0.50; 0.51; 0.52; 0.53; 0.53; 0.55; 0.55; 0.56; 0.57; 0.58; 0.58; 0.58; 0.59; 0.60; 0.60; 0.61; 0.61; 0.61; 0.62; 0.62; 0.64; 0.65; 0.66; 0.66; 0.67; 0.69; 0.70; 0.70; 0.70; 0.70; 0.72; 0.73; 0.74; 0.75; 0.75; 0.76; 0.78

Table 6: Results of E-QTSSM. Values of q, ERT and SRT at convergence, together with difference $SRT - ERT$ as percentage of ERT (a). θ_o^1 vs θ_s^1; θ_o^2 vs θ_s^2; θ_o^{nl} vs θ_s^{nl}; θ_o^p vs θ_s^p (b). Values of ERT and SRT are expressed in seconds.

	$\overline{q}$	ERT	SRT	Difference
LI–60	0.53	48627989	50476458	3.80%

(a)

	θ_o^1	θ_s^1	θ_o^2	θ_s^2	θ_o^{nl}	θ_s^{nl}	θ_o^p	θ_s^p
LI–60	0.420	0.409	0.239	0.192	0.341	0.399	0.000	0.000

(b)

To validate the approach, we have used pseudo-real data inspired from the city of Montréal and the near suburb of Laval, Québec, Canada, to produce small, medium and large instances. Numerical results confirm improvements with respect to the initial model (Bélanger et al. 2015), and small gaps between ERT and SRT. Comparing the proposed approximation methods, QTSSM outperforms PSSM in terms of both accuracy and efficiency. Moreover, to address large instances where busy fractions are not homogeneous, we propose a first attempt to deal with different busy fractions, which we have named E-QTSSM. This new yet simple approach outperforms QTSSM, producing even lower gaps between ERT and SRT and more efficient solutions with respect to BRM and QTSSM.

Future research work will be devoted to further improve the approach with different busy fractions and to include a robust optimization approach (Nicoletta et al. 2017) in the ROSA. Moreover, we want to extend the ALDP to include different priorities associated to the calls.

REFERENCES

Aboueljinane, L., E. Sahin, and Z. Jemai. 2013. "A Review on Simulation Models Applied to Emergency Medical Service Operations". *Computers & Industrial Engineering* 66(4):734–750.

Bélanger, V., E. Lanzarone, A. Ruiz, and P. Soriano. 2015. "The Ambulance Relocation and Dispatching Problem". *CIRRELT Technical Report no. 59*.

Budge, S., A. Ingolfsson, and E. Erkut. 2009. "Approximating Vehicle Dispatch Probabilities for Emergency Service Systems with Location-Specific Service Times and Multiple Units per Location". *Operations Research* 57(1):251–255.

Burman, D. Y. 1981. "Insensitivity in Queueing Systems". *Advances in Applied Probability* 13(4):846–859.

Daskin, M. S. 1983. "A Maximum Expected Covering Location Model: Formulation, Properties and Heuristic Solution". *Transportation Science* 17(1):48–70.

Fujiwara, O., T. Makjamroen, and K. K. Gupta. 1987. "Ambulance Deployment Analysis: a Case Study of Bangkok". *European Journal of Operational Research* 31(1):9–18.

Gross, D., and C. M. Harris. 2008. *Fundamentals of Queueing Theory*. John Wiley and Sons, Hoboken, New Jersey.

Harewood, S. 2002. "Emergency Ambulance Deployment in Barbados: a Multi-Objective Approach". *Journal of the Operational Research Society* 53(2):185–192.

Jarvis, J. P. 1985. "Approximating the Equilibrium Behavior of Multi-Server Loss Systems". *Management Science* 31(2):235–239.

Kergosien, Y., V. Bélanger, P. Soriano, M. Gendreau, and A. Ruiz. 2015. "A Generic and Flexible Simulation-Based Analysis Tool for EMS Management". *International Journal of Production Research* 53(24):7299–7316.

Larson, R. C. 1974. "A Hypercube Queuing Model for Facility Location and Redistricting in Urban Emergency Services". *Computers & Operations Research* 1(1):67–95.

Larson, R. C. 1975. "Approximating the Performance of Urban Emergency Service Systems". *Operations Research* 23(5):845–868.

Nicoletta, V., E. Lanzarone, V. Bélanger, and A. Ruiz. 2017. "A Cardinality-Constrained Robust Approach for the Ambulance Location and Dispatching Problem". In *Proceedings of the International Conference on Health Care Systems Engineering*, edited by P. Cappanera et al., 99–109. Springer International Publishing AG, Cham, Switzerland.

Silva, P. M. S., and L. R. Pinto. 2010. "Emergency Medical Systems Analysis by Simulation and Optimization". In *Proceedings of the 2010 Winter Simulation Conference (WSC)*, edited by B. Johansson et al., 2422–2432. Piscataway, New Jersey: IEEE.

AUTHOR BIOGRAPHIES

ETTORE LANZARONE is a Researcher at the Institute of Applied Mathematics and Information Technology (IMATI) of the National Research Council of Italy (CNR), and an Adjunct Professor at the Department of Mathematics of the Politecnico di Milano. His current research interests include robust optimization approaches and statistical methods with applications in health care, manufacturing and bioengineering. His email address is ettore.lanzarone@cnr.it.

ENRICO GALLUCCIO is currently a Data Scientist working in the private sector. His research interests include health care planning and statistics in sport. His email address is enricogalluccio@yahoo.it.

VALÉRIE BÉLANGER is an Assistant Professor at the Department of Logistics and Operations Management of the HEC Montréal, and a Member of the Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT). Her main research interests are health care logistics, emergency service management, patient and material transportation, inventory management, and network design. Her email address is valerie.3.belanger@hec.ca.

VITTORIO NICOLETTA is a Ph.D. student at the Laval University. His research interests include health care planning and statistical methods for health care. His email address is vittorio.nicoletta.1@ulaval.ca.

ANGEL RUIZ is a Professor at the Department of Operations and Decision Systems of the Laval University, a Member of the Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT), and the Head of the CIRRELT Laboratory on Healthcare Networks. His current research interests are in emergency logistics, health care management, and medical emergency planning. His email address is angel.ruiz@osd.ulaval.ca.