Impact of Ambulance Dispatch Policies on Performance of Emergency Medical Services

Cheng Siong Lim, *Member, IEEE*, Rosbi Mamat, *Member, IEEE*, and Thomas Bräunl, *Senior Member, IEEE*

***Abstract*—In ambulance location models, fleet size and ambulance location sites are two critical factors that emergency medical service (EMS) managers can control to ensure efficient delivery of the system. The ambulance relocation and dispatch policies that are studied in dynamic ambulance relocation models also significantly contribute to improving the response time of EMS. In this paper, we review dynamic ambulance relocation models from the perspective of dispatch policies. The connection between the reviewed ambulance dispatch policies and real-life policies is highlighted. Our ambulance model is based on the modified maximal covering location problem (MCLP). It is used to examine the commonly used dispatch policy and the proposed method of free-ambulance exploitation to further improve urgent call response time. Simulation results show that the proposed method can reduce the response time of urgent calls, especially during low-ambulance-supply period. We also compared the performance of EMS with and without reroute-enabled dispatch.**

***Index Terms*—Ambulance dispatch policy, ambulance location model, emergency medical services (EMSs), maximal covering location problem (MCLP).**

I. Introduction

Delivery efficiency of emergency medical services (EMS) is critical in reducing mortality and disability rates. A number of studies have found the important relationship between response time and mortality rate [1]–[4]. In real-life applications, coverage and response time are commonly used by EMS providers to evaluate delivery efficiency. A call is considered covered if it is served within a defined time threshold. Generally, an EMS response time can be defined as the interval from the time the call was received by the EMS provider to the arrival of the ambulance to the emergency scene [5]–[9].

In Montreal, QC, Canada, the implemented standard for ambulances run by "Urgences Santé" states that 90% of requests should be served within 7 min [10]. Meanwhile, the standard stated in The United States Emergency Medical Services Act is, in urban areas, 95% of requests should be served within 10 min, whereas, in rural areas, they should be served within 30 min [11]. Some countries use different response times for certain categories of calls. In the U.K., 75% of category-A calls should be served within 8 min, and 95% of category-B and -C calls should have a response time within 14 min (urban areas) and 19 min (rural areas), respectively [12].

Brotcorne *et al.* [13] classified the ambulance location models that evolved over the past 30 years into two main categories. Deterministic models were widely studied before probabilistic models emerged. Deterministic models ignore stochastic considerations but are still important in the planning stage to optimize a limited number of ambulances to provide the best services to a constituent population. For example, the maximal covering location problem (MCLP) [14] was applied by Eaton *et al.* [15] to plan the reorganization of the EMS in Austin, TX. As a result, $3.4 million in construction costs and $1.2 million annually in operating costs were saved in 1984. Furthermore, the average response time was reduced, even with an increase in calls for the services.

The location set covering model (LSCM) [16] was one of the earliest models introduced for seeking a minimum number of ambulances to cover all demands. The strategic location sites that provide full coverage with minimum ambulances can be identified from a given set of potential ambulance location sites. In this model, the number of ambulances is unlimited, and a demand node is assumed to be covered if it can be reached within a time threshold. LSCM can be used to determine the right number of ambulances and strategic location sites to cover all demands.

The later models such as MCLP [14], modified MCLP [17], the double standard model (DSM) [18], the tandem equipment allocation model (TEAM) [19], and the facility-location, equipment-emplacement technique (FLEET) [19] are coverage maximization models that optimize the demand coverage with a limited number of ambulances. In MCLP, coverage may become an issue when ambulances become busy. Hence, the modified MCLP [17] takes multiple coverage optimization into account without increasing the total number of ambulances. This improves the backup coverage when ambulances become busy. DSM further extends the multiple coverage optimization to include double covering constraints. The dynamic DSM (DDSM^t) [10] is a dynamic model of DSM that dispatches and relocates ambulances with practical considerations at each instant at which a call is registered. Recognizing the fact that, on occasions when heterogeneous vehicles may be dispatched to an emergency scene, TEAM and FLEET maximize the coverage of two vehicle types.

On the other hand, probabilistic models such as the maximum expected covering location problem formulation [20] and the maximum availability location problem [21] take into account the probability that an ambulance becomes busy. In general, both deterministic and probabilistic models are useful planning tools for determining fleet size and strategic location sites based on the objective function of the respective models. However, both models lack the capability to handle the fluctuating demand over time. This is overcome by dynamic models that periodically update ambulance positions throughout the day [13].

In the next section, the details of the dispatch policies used in dynamic models and real-life applications are reviewed. We present our ambulance location model in Section III. Simulation setup and results for commonly used dispatch policies are reported in Sections IV and V. Finally, the discussion and conclusion of this work are presented in Sections VI and VII.

II. Ambulance Dispatch Policy Reviews

A. Related Simulation Work

There has been limited research work carried out for dynamic models in the past three decades at the time of reviewing by Brotcorne *et al.* [13]. In this section, five dispatch policies used in the dynamic models [9], [10], [22], [23] are identified and presented. Ambulance dispatch is the process of assigning an ambulance to answer an emergency call.

There are two dispatch policies evaluated by Repede and Bernardo [9]. The first dispatch policy always assigns the closest available ambulance to a call scene. Hence, there is the possibility that the closest ambulance is unable to respond within the time threshold. If

Manuscript received October 13, 2009; revised July 13, 2010 and November 3, 2010; accepted November 28, 2010. Date of publication February 4, 2011; date of current version June 6, 2011. This work was supported by Universiti Teknologi Malaysia. The Associate Editor for this paper was M. Chowdhury.

C. S. Lim and R. Mamat are with the Universiti Teknologi Malaysia, 81310 Skudai, Johor Darul Ta'zim, Malaysia (e-mail: limchengsiong@fke.utm.my; rosbi@fke.utm.my).

T. Bräunl is with the Centre for Intelligent Information Processing Systems, School of Electrical, Electronic, and Computer Engineering, The University of Western Australia, Perth, WA 6009, Australia (e-mail: ciips@ee.uwa.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2010.2101063

this happens and, coincidently, the next immediate call scene is from the currently dispatched ambulance's district and there is no other ambulance that can respond within the time threshold, then the closest ambulance from another district is dispatched. As a result, both call scenes are unable to be served within the time threshold.

In view of this, Repede and Bernardo [9] proposed an alternative dispatch policy. If no ambulance can respond within the time threshold, then dispatch the ambulance with the least likelihood of receiving a call in its primary district. On the other hand, if one or more ambulances can respond within the time threshold, then any of these can be dispatched. The simulated result showed a small coverage improvement at the expense of a slight increase in mean response time for the alternative dispatch policy. For example, at a fleet size of four ambulances, the mean coverage was slightly improved from 0.6247 to 0.6309 by using the alternative dispatch policy. However, the mean response time slightly increased from 12.22 to 12.23 min.

In another work by Gendreau *et al.* [10], calls are served in decreasing order of priority. The closest available ambulance is dispatched to serve a call. For urgent calls, the ambulances en route to new location sites are also included in the dispatch consideration. Moreover, an ambulance already assigned to a less urgent call can be reassigned to an urgent call if the ambulance is the closest to the urgent-call scene, and there is an alternative ambulance capable of covering the less urgent call within the remaining time. In cases without an ambulance that is capable of covering the urgent call, the closest ambulance already assigned to a less urgent call can be reassigned.

In the research of [22], the calls are categorized as PRIO 1, 2, and 3 based on the degree of urgency. PRIO-1 calls are the most urgent and life-threatening calls. The closest available ambulance is always dispatched to answer a PRIO-1 call. The ambulance assignment for PRIO-2 and -3 calls is based on the preparedness impact due to the assignment of ambulances that can respond within the time threshold. Preparedness is the ability to, within a reasonable time, offer EMS to the inhabitants in a specific geographical area [24]. The ambulance with the lowest impact to the preparedness of all zones is dispatched. On the contrary, if there is no ambulance that can answer PRIO-2 and -3 calls within the time threshold, then the closest ambulance is dispatched. During this ambulance assignment process, any ambulances on their way to less urgent calls are also considered for more urgent call assignments. For example, an ambulance on its way to a PRIO-3-call scene can be reassigned to a new PRIO-2 or -1 call. It is interesting to note that the authors have introduced pseudopriorities in the EMS system. The pseudopriority for a less urgent call, e.g., PRIO 3, that has been placed in the waiting queue for a certain time can be changed to PRIO 2 to shorten the waiting period.

Maxwell *et al.* [23] used the simplest dispatch policy. Calls are served in decreasing order of priority. For calls with the same priority levels, first-in–first-out order is applied. The closest available ambulance is dispatched to serve a call. If there is no available ambulance, the call is placed into a waiting queue. No reassignment is allowed for an ambulance already assigned from a less urgent call to an urgent call, even if the ambulance is the closest to the urgent-call scene.

B. Dispatch Policies in Real-Life Applications

Some EMS systems use police or firefighters as first responders, whereas others solely rely on ambulances. In general, ambulances can be divided into basic life support (BLS) and advanced life support (ALS) ambulances. Some EMS systems have both types (two tier), whereas others have only a single type (one tier) [25]. The earliest and simplest method is to dispatch the ambulance based on the order of received calls, regardless of the calls' urgency [12]. (We name it first-in–first-out dispatch.) With the introduction of the priority dispatch, the received calls are first prioritized before dispatching BLS or ALS ambulances based on the calls' urgency. Note that we limit the scope of "priority dispatch" to the process of prioritizing the incoming emergency calls.

Criteria-based dispatch (CBD) [26] and advanced medical priority dispatch (AMPD) [27] are considered as priority dispatch [28]. The CBD system is based on predetermined guidelines to help the dispatcher in reaching a priority decision, whereas the AMPD system relies on scripted questions and protocols in the process of prioritizing a call. Priority dispatch is now used by many EMS providers, especially in developed countries [12], [29]. It is not suitable for resource-limited EMS providers as extra resources are required to implement priority dispatch [29]. In the U.K., first-in–first-out dispatch was replaced with priority (AMPD) dispatch in April 2001. The percentage of category-A calls served within 8 min has increased from 70.7% (2001–2002, with about three months using first-in–first-out dispatch) to 74.6% (2002–2003, with priority dispatch) [5].

After a call has been prioritized using priority dispatch or processed using first-in–first-out dispatch, the most suited ambulance has to be identified and dispatched. The closest dispatch is the most commonly used method for ambulance assignment [9]. For a one-tier EMS system, either the closest available BLS or ALS ambulances are dispatched. We call it closest (uniform) dispatch. For a two-tier EMS system using priority dispatch, BLS ambulances are dispatched for most of the emergency calls. ALS ambulances are spared and dispatched only for high-priority calls. We name this closest (tiered) dispatch.

Studies show that an EMS system using a tiered system performs better than a one-tier all-ALS system [30]. Generally, a two-tier EMS system has a higher cardiac arrest survival rate than a one-tier EMS system [25], [31]. In addition, two-tier EMS systems provide more cost savings due to the increased ALS ambulance utilization. The accuracy in prioritizing is very important to ensure that ALS ambulances can be safely excluded from lower priority calls [32]. On the contrary, the accuracy of prioritizing is not an issue for a one-tier all-ALS system as ALS ambulances are always dispatched.

As the communication technology advances, mobile data terminal and standard radio communication are now commonly installed on the ambulances. It is then possible to reroute an ambulance dispatched from lower to higher priority call [6]. (We name it reroute-enabled dispatch.) In addition, communication technology also enables the latest condition of a victim to be updated from time to time while the ambulance is on the road. Thus, the priority of the call can be upgraded or downgraded accordingly. The ambulance en route to the emergency scene can also be canceled [33]. (We name it priority-update-enabled dispatch.) Some EMS systems provide pre-arrival instructions prior to ambulance arrival [6], [28]. The instructions are medically approved and provided by a call taker in the ambulance dispatch center to the caller to provide an aid to the victim.

With the advent of a geographic information system (GIS) and information technology, ambulances in some developed countries are equipped with global positioning system [34]. Ambulances can be managed in a more efficient manner to provide faster response time through a computer-aided dispatch (CAD) system that can trace the dynamic status and location of ambulances.

C. Dispatch Policy Decomposition

There are many processes involved in a real-life ambulance dispatch system, starting from receipt of an emergency call until the dispatch of an ambulance to the emergency scene. Detailed EMS processes are presented and discussed in [4], [6], [9], [35], and [36]. Based on our analysis on the mentioned dispatch policies and real-life ambulance dispatch processes, an ambulance dispatch policy must include the method of call queuing and the way of assigning an ambulance to answer an emergency call in the queue. This information can be

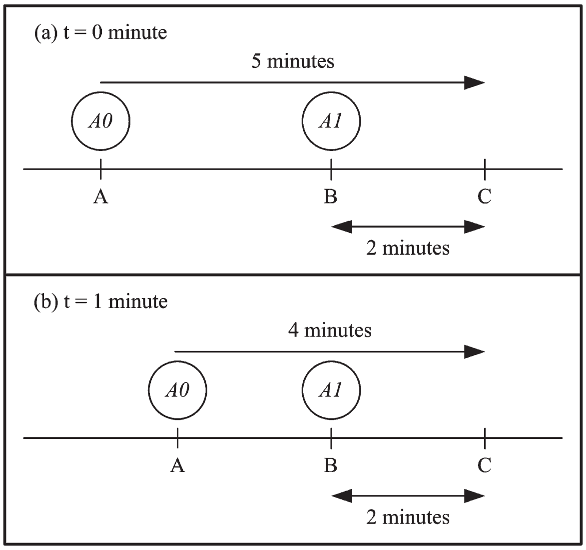

Fig. 1. Latest position of the ambulances relative to the call scenes at (top) $t = 0$ and (bottom) $t = 1$.

regarded as the core of the ambulance dispatch policy, which summarizes the complicated processes involved in a real-life ambulance dispatch system. Add-on dispatch is a supplementary method used in ambulance dispatch policy to achieve a specific objective.

Priority and first-in–first-out dispatches are core dispatches used in sorting the calls in the waiting queue to be served by ambulances. Meanwhile, the closest dispatch is the core dispatch used in assigning a proper ambulance to answer the call in the queue. Other nonclosest dispatches combine coverage with probability [9] or preparedness [22] to determine the proper ambulance for call assignment. However, there may be legal complications for not dispatching the closest available unit [9].

Reroute-enabled dispatch [6] is an add-on dispatch to exploit the active ambulance fleet to further improve the response time of urgent calls. The fallout of the dispatch is the increase in waiting time of lower priority calls. Pseudopriority introduced in [22] can upgrade the call from lower to higher priority, so that the lower priority call with long waiting time can be shortened. Priority-update-enabled dispatch and the use of pre-arrival instructions are categorized as add-on dispatches.

Free-ambulance-exploitation dispatch refers to our proposed add-on dispatch that enables free-ambulance exploitation to improve the response time of an urgent call. Among the dispatch policies reviewed in Section II-A, only [10] and [22] permit reassignment of an ambulance from a less urgent call to an urgent call. In other words, the less urgent call dispatch plan, which consists of ambulance and call assignment, can be modified, so that the best ambulance can be assigned to an urgent call. Nevertheless, none of the dynamic models [9], [10], [22], [23], [37], [38] has looked into the possibility of reassigning an ambulance that just completed a call to further improve the urgent call dispatch plan. In this paper, we analyze its impact on the improvement of the response time of urgent calls.

Assume that there are two ambulances $A0$ and $A1$ at location sites A and B (see Fig. 1). Ambulance $A1$ is busy serving a call when an urgent call for scene C is received. Thus, the only free ambulance $A0$ is dispatched. After 1 min, ambulance $A1$ completes its current call and becomes free. Under the current dispatch policies [9], [10], [22], [23], ambulance $A1$ has no contribution to the current urgent call dispatch plan. Our proposed method of free-ambulance exploitation improves the dispatch plan by reassigning ambulance $A1$ to answer the call for scene C. As a result, the response time of the urgent call is reduced from 5 to 3 min.

Ambulance reroute is permitted in real-life EMS [6] and being used in simulation in [10] and [22]. The performances of EMS with priority (AMPD) dispatch against first-in–first-out dispatch [5], closest dispatch versus nonclosest dispatch [9], and two-tier against one-tier EMS systems [25], [30]–[32] have been studied under respective conditions. Although reroute-enabled dispatch is commonly used, no specific study on its performance has been reported. Thus, the performance of EMS with and without reroute-enabled dispatch is also included in our simulation.

Finally, all the mentioned dispatch methods are properly organized in Table I. Note that priority (simplified) dispatch is added to differentiate between the simplified version of priority dispatch used in simulations with real-life CBD and AMPD. Comparisons of the urgent and less urgent dispatch policies mentioned in Section II-A are given in Tables II and III, respectively.

III. Ambulance Location Model

Our ambulance model is based on MCLP [14]. The model is defined as a graph $G = (V \cup W, E)$, where V and W are two vertex sets representing demand points and potential ambulance location sites, respectively. E is an edge set representing the distance between two vertices. We add the requirement that α proportion of the demand must lie within time threshold r. The demand at point $i \in V$ (which is denoted by d_i) is said to be covered by location $j \in W$ if and only if $t_{ij} < r$, where t_{ij} is the shortest travel time from location j to point i. The set of ambulance location sites covering demand point i is denoted by $W_i \subseteq W$, and the number of required ambulances to achieve α proportion of the demand coverage is denoted by p. Meanwhile, $P_{\max}$ and p_{extra} denote the number of available and extra ambulances, respectively. The binary variable x_j is equal to 1 if and only if an ambulance is placed at location j. On the other hand, y_i is equal to 1 if and only if point i is covered by at least one ambulance within r. The extended MCLP model is written as follows: It is important to note that (3, 5, 6) are introduced and added to the original MCLP

$$\text{Maximize} \sum_{i \in V} d_i y_i \tag{1}$$

Subject to

$$\sum_{j \in W_i} x_j \geq y_i \qquad \forall i \in V \tag{2}$$

$$\sum_{i \in V} d_i y_i \geq \alpha \sum_{i \in V} d_i \tag{3}$$

$$\sum_{j \in W} x_j = p \tag{4}$$

$$P_{\max} \geq p \tag{5}$$

$$p_{\text{extra}} = P_{\max} - p \tag{6}$$

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

$$y_i \in \{0, 1\}, \qquad i \in V. \tag{8}$$

The objective (1) is to maximize the demand coverage. Constraint (2) counts the number of ambulances that cover each demand point. Constraint (3) expresses the coverage requirement that α proportion of the demand must be covered. Constraint (4) counts the total number of required ambulances for maximized demand coverage. Constraint (5) ensures that the total number of required ambulances in constraint (4) can never exceed the total number of deployed ambulances at dispatch centers. Constraint (6) counts the extra ambulances that can be used for multiple coverage but in conjunction with constraints (4) and (5).

The main purpose of this paper is to determine the effectiveness of the proposed free-ambulance-exploitation dispatch using a real-life

TABLE I
COMPONENTS OF AMBULANCE DISPATCH POLICY

Core Dispatches		Add-on Dispatches
Calls Queuing Methods	Ambulance Assignment Methods	
First in first out dispatch, priority (CBD) dispatch, priority (AMPD) dispatch, priority (simplified) dispatch	Closest (uniform) dispatch, closest (tiered) dispatch, non-closest (coverage combines with probability) dispatch, non-closest (coverage combines with preparedness) dispatch	Reroute enabled dispatch, priority update enabled dispatch, pre-arrival instructions, pseudo priority, GIS support, free ambulance exploitation dispatch

TABLE II
COMPARISON OF THE URGENT DISPATCH POLICIES

Models / References	Core Dispatches		Add-on Dispatches
	Calls Queuing Methods	Ambulance Assignment Methods	
TIMEXCLP [9] - Closest dispatch	First in first out dispatch	Closest (uniform) dispatch	None
TIMEXCLP [9] - Alternative dispatch	First in first out dispatch	Non-closest (coverage combines with probability) dispatch	None
RP^t [10]	Priority (simplified) dispatch	Closest (uniform) dispatch	Reroute enabled dispatch
DYNAROC [22]	First in first out dispatch	Closest (uniform) dispatch	Reroute enabled dispatch
By Maxwell *et al.* [23]	Priority (simplified) dispatch	Closest (uniform) dispatch	None

TABLE III
COMPARISON OF THE LESS-URGENT DISPATCH POLICIES

Models / References	Core Dispatches		Add-on Dispatches
	Calls Queuing Methods	Ambulance Assignment Methods	
TIMEXCLP [9] - Closest dispatch	First in first out dispatch	Closest (uniform) dispatch	None
TIMEXCLP [9] - Alternative dispatch	First in first out dispatch	Non-closest (coverage combines with probability) dispatch	None
RP^t [10]	Priority (simplified) dispatch	Closest (uniform) dispatch	None
DYNAROC [22]	First in first out dispatch	Non-closest (coverage combines with preparedness) dispatch	Pseudo priority
By Maxwell *et al.* [23]	Priority (simplified) dispatch	Closest (uniform) dispatch	None

proven dispatch policy and ambulance location model. This paper provides a better insight into a real-life proven policy and model that can be easily adopted with performance assurance at minimal cost. Our ambulance location model is an extension of MCLP [14], which has been successfully implemented in Austin, TX [15]. The most commonly used dispatch methods in real-life EMS, priority, and closest dispatches [9], [12] are adopted for EMS simulation. It has not been our intention in this paper to compare the impact of other ambulance dispatch policies or ambulance location models that have yet been proven successful in real-life EMS systems.

IV. SIMULATION SETUP

The method of using a grid to partition the EMS annual demand into zones has been implemented in [22] and [38] for coverage study. We define a reasonable hypothetical region of 4096 km^2 (64 km $\times$ 64 km) that is divided into 256 zones. This dimension is assumed to represent a large-scale problem [38]. Fig. 2 shows the defined spatial distribution of the EMS demand in the hypothetical region. ($i \in V$ and d_i are provided). Assume that there are 22 potential ambulance location sites in the hypothetical region (shown in Fig. 3) where ambulances can be strategically placed. ($j \in W$ is given.) A call is considered covered if served within 10 min ($r = 10$ min). We target to have a minimum of 0.8 proportion of the demand coverage ($\alpha = 0.8$).

Based on our defined ambulance location model, a simple full-combination search is performed to identify the strategic ambulance location sites. The identified sites are as shown in Fig. 4. In other words, we need a minimum of eight ambulances ($p = 8$) to provide basic coverage for the hypothetical region with $\alpha = 0.8$. The total number of deployed ambulances $P_{\max}$ ranging from eight to 16 is used in the simulation. The extra ambulances p_{extra} are positioned at the strategic ambulance location sites to provide backup coverage.

Calls are categorized into urgent and less urgent calls with a ratio of 1 : 1. An urgent call has higher priority than a less urgent call. The real-life data provided by Urgences Santé have a higher ratio of urgent calls [10]. On the contrary, the data collected in the U.K. show a higher proportion of less urgent calls [5]. Many factors can cause the deviation, such as the method and accuracy of prioritizing the calls [39]–[42]. Thus, we choose the 1 : 1 call ratio as our reference to ensure that the performance of the evaluated dispatch policies is not biased by the different proportions of call types.

Based on the spatial distribution, five sets of incoming calls are randomly generated for simulation. Each set contains 40 predefined incoming calls for an 8-h simulation time. Different random seeds are

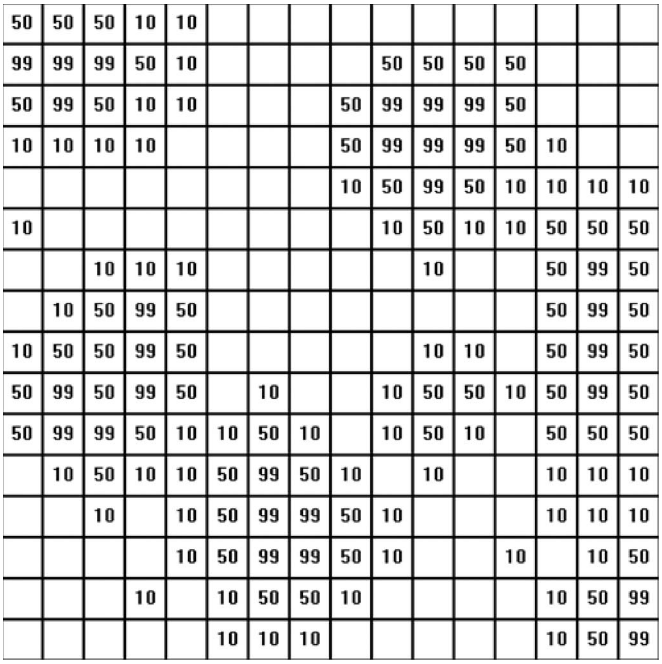

Fig. 2. Annual demand for EMS in the hypothetical region normalized to 0–99.

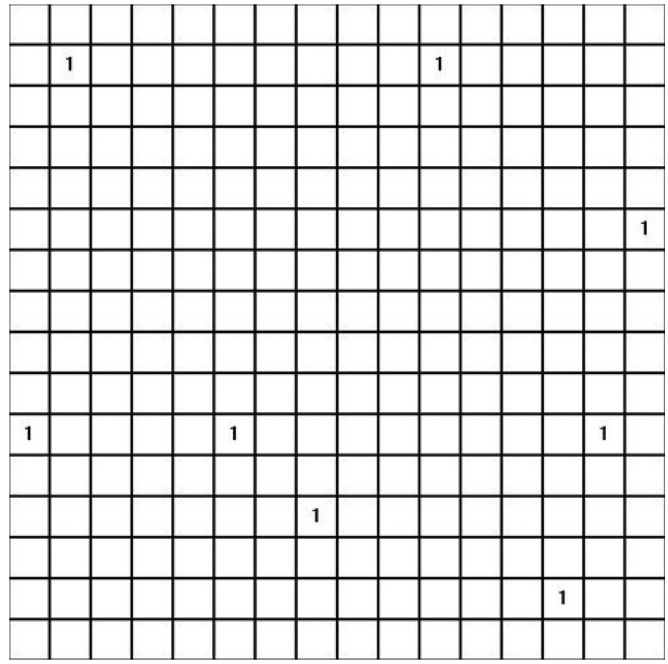

Fig. 4. Identified strategic ambulance location sites that provide a minimum 0.8 proportion of the demand coverage.

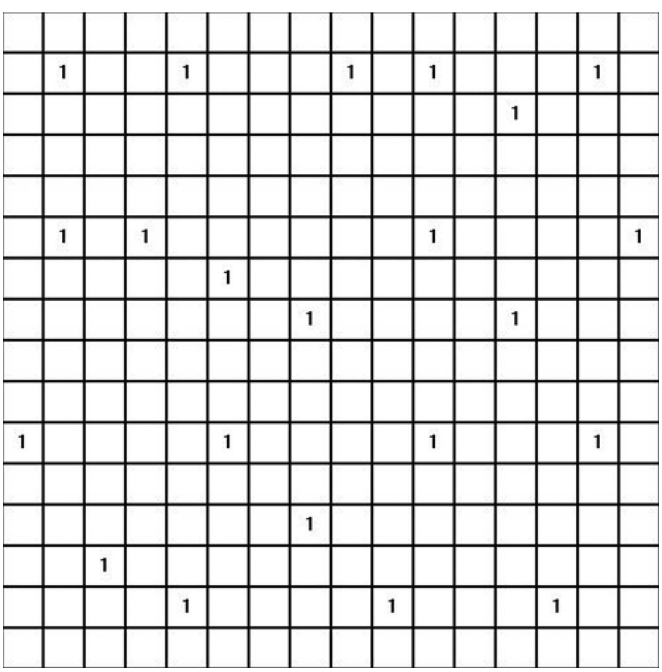

Fig. 3. Location of potential ambulance location sites marked with "1."

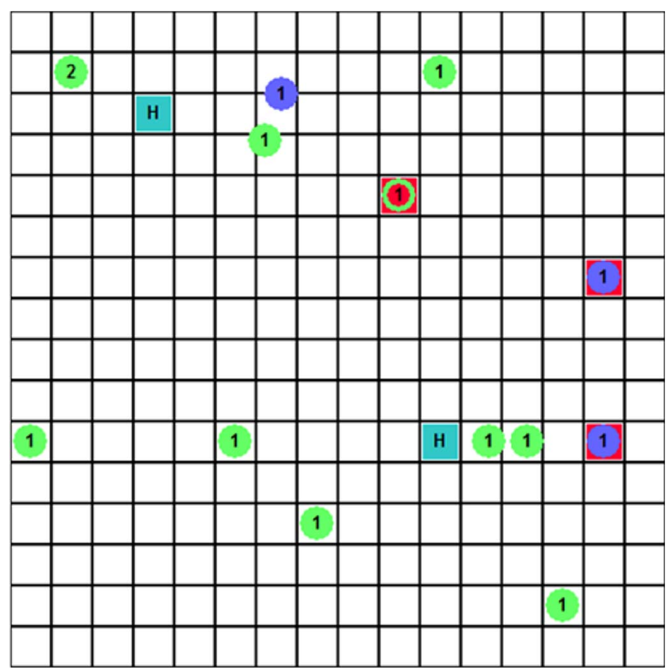

Fig. 5. Graphical representation for the delivery of EMS during simulation. Circles represent the ambulances, and the numbers inside tell the total number of ambulances at a specific location. The color of the circle is used to indicate the status of an ambulance: Green is free; yellow and red mean less urgent and urgent call dispatch, respectively; and blue stands for busy at serving or transporting a victim. Yellow and red are used to indicate the type of call that originated from a zone, i.e., less urgent or urgent call.

used for each set generation. The zone with a higher number of annual demand has higher probability to be assigned with calls. Thus, each set contains 40 calls with very different combinations of zones. As the ratio of urgent to less urgent calls is 1 : 1, thus, by using different random seeds, 50% of the calls are randomly picked to become urgent calls, whereas the rest are less urgent. Next, different random seeds are used to randomize the time of call events over the 8-h period. Thus, the generated set of calls is unique, with an average call rate of five calls per hour, reflecting the spatial distribution of the demand with 1 : 1 ratio of urgent to less urgent calls.

Other relevant information is also defined. Ambulance speed is set at 60 km/h. Each call requires one ambulance, and a fixed 10 min is used for both times at the scene and at the hospital. All patients need to be transported to a hospital. There are two hospitals located in the hypothetical region. Zero turn-out time is used for free-ambulance redeployment. The fixed ambulance speed and interval at the scene

TABLE IV
DISPATCH METHODS USED IN SIMULATION SETTINGS

Settings	Core Dispatches		Add-on Dispatches	
	Calls Queuing Methods	Ambulance Assignment Methods	Reroute Enabled Dispatch	Free Ambulance Exploitation Dispatch
Setting R	Priority (simplified) dispatch	Closest (uniform) dispatch	Enabled (Urgent calls only)	Disabled
Setting RF	Priority (simplified) dispatch	Closest (uniform) dispatch	Enabled (Urgent calls only)	Enabled (Urgent calls only)
Setting B	Priority (simplified) dispatch	Closest (uniform) dispatch	Disabled	Disabled
Setting F	Priority (simplified) dispatch	Closest (uniform) dispatch	Disabled	Enabled (Urgent calls only)

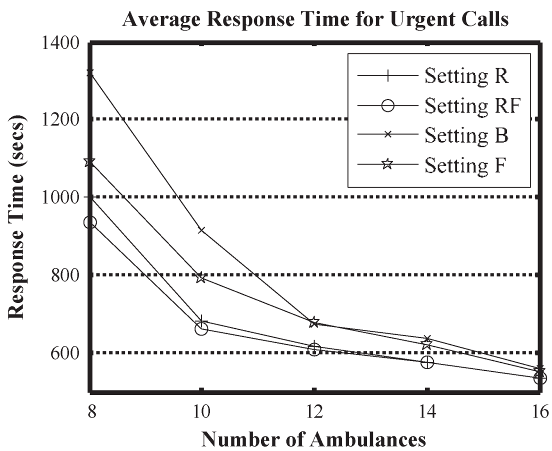

Fig. 6. Average response time for urgent calls at different supplies of ambulances. At eight ambulances, Setting F shows a significant response time reduction of 17.32% in comparison with Setting B.

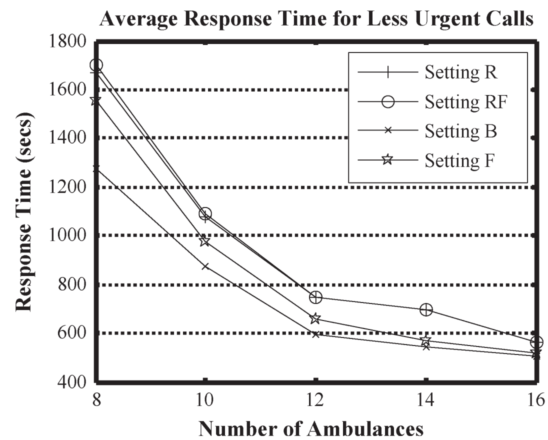

Fig. 7. Average response time for less urgent calls at different supplies of ambulances. Higher average response time is obtained for the setting with better urgent response time optimization.

and the hospital can be deemed as mean values, which are within the range of or close to the actual data collected in [7] and [10]. From this information, the time for a complete ambulance cycle can be calculated.

Since the objective of our simulation is to examine the relative performance of various ambulance dispatch policies, Euclidean distance is used to compute the distances among hospitals, ambulances, and call scenes. This method is widely used in a vehicle-routing problem [43]. Simulation using an advanced method for the best path identification [44], [45] and an actual map with real traffic and streets can be more complicated and suitable for the implementation stage to simulate the targeted real-life problem. These will be considered in our future work.

The proposed method of free-ambulance exploitation, ambulance location model, dispatch policies, and the described ambulance dynamic information are coded in C++. The dynamic of the simulation can be observed as shown in Fig. 5.

V. SIMULATION RESULTS

There are four settings being tested in the simulation, which are as defined in Table IV. Both reroute-enabled and free-ambulance exploitation dispatches are applied to optimize urgent calls. Each setting is simulated with five sets of randomly generated data. The numbers of ambulances used in the simulation are eight, 10, 12, 14, and 16. Thus, there are a total of 100 simulations being performed.

Fig. 6 shows that the settings with free-ambulance-exploitation dispatch (Settings RF and F) can reduce the average response time for urgent calls. During the low-ambulance-supply period, the impact is more significant. At eight ambulances, the average response time of Setting RF is 6.44% lower than that of Setting R. Setting F shows a 17.32% reduction in comparison to Setting B. As the number of ambulances increases, the impact of both reroute-enabled and free-ambulance exploitation dispatches decreases. This is because, with a larger fleet size, more free ambulances are available to answer a newly incoming urgent call. Thus, it reduces the number of urgent calls that both add-on dispatches can optimize.

For comparison of urgent call optimization among different ambulance dispatch policies, the plotted graph of Setting B is used as our reference since it adopts only the core dispatches. Settings R, RF, and F show that urgent calls can be optimized through adoption of proper add-on dispatches. The graphs show that Setting R outperforms Setting F. This is because reroute-enabled dispatch can better optimize urgent call response time than free-ambulance-exploitation dispatch. The use of both add-on dispatches (Setting RF) can outperform other settings.

Fig. 7 shows that the settings that provide better urgent call response time in Fig. 6 perform contrary for less urgent calls. Ambulances can be treated as a limited resource in EMS. The dispatch optimized for urgent calls can lead to the lack of available units to answer less urgent calls. Nevertheless, the average response time for all calls is not much affected (see Fig. 8) by the dispatch that optimizes urgent call response time. Note that pseudopriority is not used in our simulation to reduce the response time of less urgent calls with a long waiting time. This is because pseudopriority can increase the average urgent call response time and thus can indirectly affect the performance of the examined dispatch policies. However, it can be included in real-life applications.

The three graphs shown in Figs. 6–8 demonstrate the similar trend where the response time can be further reduced by increasing the number of ambulances. The effect of ambulance increment is lower at a larger fleet size. At double the basic required number of ambulances to provide basic coverage, the average response times of urgent and

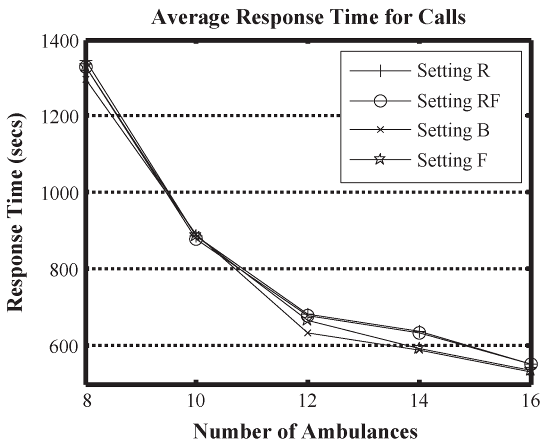

Fig. 8. Average response time for all calls at different supplies of ambulances. There is no obvious improvement achieved by any dispatch policy.

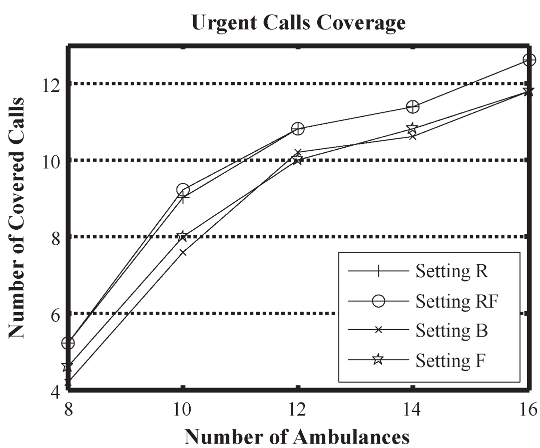

Fig. 9. Urgent call coverage at different supplies of ambulances. Better urgent coverage achieved by Settings R and RF, which permit the reconsideration of ambulances assigned to less urgent calls to urgent calls.

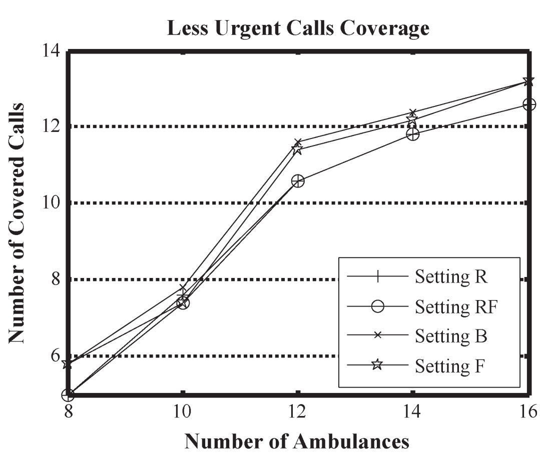

Fig. 10. Less urgent call coverage at different supplies of ambulances.

less urgent calls, regardless of the settings, are converging to a value between 500 and 600 s.

Figs. 9–11 show the performance of the EMS as in the number of calls being covered. From Fig. 9, the simulations using Settings R and RF clearly outperform the simulations using Settings B and F. The contrary performance is obtained for less urgent calls (see Fig. 10).

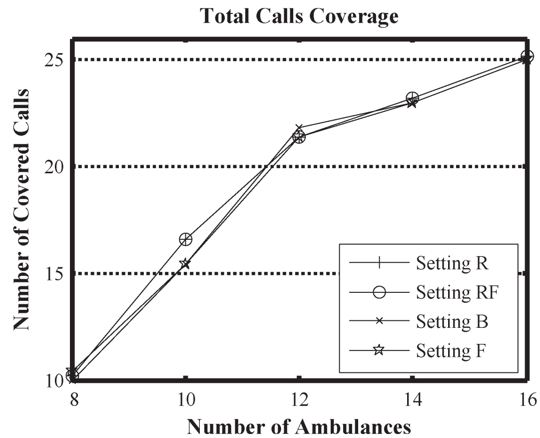

Fig. 11. Total call coverage at different supplies of ambulances. There is no significant difference in total call coverage among settings quit different optimization methods.

There is no significant effect on the total call coverage (see Fig. 11) due to diverting ambulances for urgent call response time optimization.

VI. Discussion

From Table I, an ambulance dispatch policy can be formed using various dispatch methods. In our opinion, a complete ambulance dispatch policy evaluated in EMS simulation should clearly inform the readers of its core and add-on (if any) dispatches. The readers can comprehend and trace it back to real-life ambulance dispatch policies and processes. Furthermore, this is very important for other researchers as the adoption of different ambulance dispatch policies can yield different EMS performances [5], [9], [25], [30]–[32].

From our simulation results and the findings by other researchers [5], [9], [25], [30]–[32], we summarize the advantages and disadvantages of various dispatch methods in Table V for general reference. There is no single ambulance dispatch policy that fits all. The adoption of any ambulance dispatch policy has to fulfill the objectives and performance defined by respective EMS providers within the constraint of available funding and other resources.

The survey shows that EMS has quite differently developed in different countries due to variations in funding [40]. Developed countries with high income can afford to implement CAD systems for better ambulance management [34], [35]. On the other hand, developing countries with poorly developed EMS are facing a critical financial problem even to implement a comprehensive communication network between hospitals and ambulances [46], [47]. There is also a lack of telecommunication infrastructure for the communities to quickly contact a hospital [48]. In addition, the upgrading process involved in an ambulance dispatch system can be very complicated and costly [35], [49]. In many countries, other healthcare issues may take priority in funding allocation.

On the other hand, the obtained results can be used to estimate the performance of EMS when there is a change in the factors related to the EMS process cycle. For example, the road network improvement and the efficiency enhancement on dispatch processes have a direct impact on the EMS response time. As a result of reduction in average call response time, ambulances are getting more idle time. The probability of ambulances that are free to answer a newly incoming call is higher. The effect is almost similar to an increasing fleet size, as demonstrated in our simulation, where ambulances are getting more idle time.

TABLE V
ADVANTAGES AND DISADVANTAGES OF VARIOUS DISPATCHES

Variance of Dispatches	Advantages	Disadvantages
First in first out dispatch	• The simplest calls sorting method • Requires minimum resources allocation in the ambulance dispatch centre for calls processing	• Lacking of optimization for higher priority calls
Priority dispatch	• Most commonly used calls sorting method • Optimization for higher priority calls	• Can affect the response time of lower priority calls • Requires calls resorting when a new call is added in the waiting queue (Relative to first in first out dispatch)
Closest dispatch	• Most commonly used ambulance assignment method where the best available unit is dispatched to answer the emergency call with the fastest response time • Can be exploited to optimize higher priority calls by using closest (tiered) dispatch after calls prioritizing	• Requires mobile data terminal or standard radio communication to be installed on the ambulances if the ambulances are placed at strategic location sites
Non-closest (coverage combines with probability or preparedness) dispatch	• Ambulances are reserved for expected future calls and thus can maximize EMS coverage	• May cause legal issues for not dispatching the closest available ambulance • Requires more complicated probability or preparedness evaluation (Relative to closest dispatch) • Requires mobile data terminal or standard radio communication to be installed on the ambulances if the ambulances are placed at strategic location sites
Reroute enabled dispatch	• Optimization for higher priority calls	• Higher ambulance travelling cost due to rerouting • Can affect the response time of lower priority calls • Disturbance to ambulance crew • Requires mobile data terminal or standard radio communication to be installed on the ambulances • Requires extra resources to implement add-on dispatch
Priority update enabled dispatch	• Increases the efficiency of ambulances utilization as the status of a call can be upgraded, downgraded or cancelled	• Disturbance to ambulance crew • Requires mobile data terminal or standard radio communication to be installed on the ambulances • Requires extra resources to implement add-on dispatch
Pre-arrival instructions	• To provide an aid to the victim prior to ambulance arrival and thus can reduces mortality rate	• Requires extra resources to implement add-on dispatch
Pseudo priority	• The lower priority calls with long waiting time can be shortened	• Can affect the response time of higher priority calls • Requires extra resources to implement add-on dispatch
GIS support	• Increases the efficiency of ambulances utilization through instant status and location update	• Requires extra resources to implement add-on dispatch
Free ambulance exploitation dispatch	• Optimization for higher priority calls	• Can affect the response time of lower priority calls • Disturbance to ambulance crew • Requires mobile data terminal or standard radio communication to be installed on the ambulances • Requires extra resources to implement add-on dispatch

VII. CONCLUSION

The performance of EMS can help in reducing the mortality rate. It is less emphasized in developing countries, despite the improvement being more necessary. Even though a larger fleet size can improve the EMS performance, the efficiency of resource utilization will decrease [7]. Our important finding in this paper is that the response time of urgent calls can be reduced through adoption of an appropriate dispatch policy without increasing the fleet size.

From the simulation results, reroute-enabled dispatch can significantly improve the response time of urgent calls. The proposed method of free-ambulance exploitation can further enhance the performance of EMS in urgent calls, especially during the low-ambulance-supply period. The proposed method is computationally proven to work for the commonly used ambulance dispatch policy. The obtained results also show that the process of diverting ambulances has a low impact on the total call coverage and the average response time for all calls. However, urgent call optimization using both dispatches can lead to lower performance for less urgent calls.

The results presented in this paper are mainly based on the EMS simulation in the hypothetical region. A more realistic network with actual traffic and streets of varying sizes and capacities are to be built into the current simulator. Actual demand patterns will be considered in the formation of the simulation data. It is expected that we will be able to evaluate the EMS performance closer to the actual implementation. We will report the findings in forthcoming papers.

REFERENCES

[1] R. Sánchez-Mangas, A. García-Ferrrer, A. de Juan, and A. M. Arroyo, "The probability of death in road traffic accidents. How important is a quick medical response?" *Accid. Anal. Prev.*, vol. 42, no. 4, pp. 1048–1056, Jul. 2010.

[2] R. P. Gonzalez, G. R. Cummings, H. A. Phelan, M. S. Mulekar, and C. B. Rodning, "Does increased emergency medical services prehospital time affect patient mortality in rural motor vehicle crashes? A statewide analysis," *Amer. J. Surg.*, vol. 197, no. 1, pp. 30–34, Jan. 2009.

[3] R. B. Vukmir, "Survival from prehospital cardiac arrest is critically dependent upon response time," *Resuscitation*, vol. 69, no. 2, pp. 229–234, May 2006.

[4] "Part 12: From science to survival: Strengthening the chain of survival in every community," *Resuscitation*, vol. 46, no. 1–3, pp. 417–430, Aug. 2000.

[5] U.K. National Statistics, Ambulance Services England 2008–2009, NHS Inform. Cent., 2009.

[6] M. Castrén, R. Karlsten, F. Lippert, E. F. Christensen, E. Bovim, A. M. Kvam, I. Robertson-Steel, J. Overton, T. Kraft, L. Engerstrom, and L. Garcia-Castrill Riego, "Recommended guidelines for reporting on emergency medical dispatch when conducting research in emergency medicine: The Utstein style," *Resuscitation*, vol. 79, no. 2, pp. 193–197, Nov. 2008.

[7] P. T. Pons and V. J. Markovchick, "Eight minutes or less: Does the ambulance response time guideline impact trauma patient outcome?" *J. Emerg. Med.*, vol. 23, no. 1, pp. 43–48, Jul. 2002.

[8] A. K. Marsden, "Getting the right ambulance to the right patient at the right time," *Accid. Emerg. Nurs.*, vol. 3, no. 4, pp. 177–183, Oct. 1995.

[9] J. F. Repede and J. J. Bernardo, "Developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky," *Eur. J. Oper. Res.*, vol. 75, no. 3, pp. 567–581, Jun. 1994.

[10] M. Gendreau, G. Laporte, and F. Semet, "A dynamic model and parallel tabu search heuristic for real-time ambulance relocation," *Parallel Comput.*, vol. 27, no. 12, pp. 1641–1653, Nov. 2001.

[11] M. O. Ball and L. F. Lin, "A reliability model applied to emergency service vehicle location," *Oper. Res.*, vol. 41, no. 1, pp. 18–36, Jan./Feb. 1993.

[12] J. J. M. Black and G. D. Davies, "International EMS Systems: United Kingdom," *Resuscitation*, vol. 64, no. 1, pp. 21–29, Jan. 2005.

[13] L. Brotcorne, G. Laporte, and F. Semet, "Ambulance location and relocation models," *Eur. J. Oper. Res.*, vol. 147, no. 3, pp. 451–463, Jun. 2003.

[14] R. L. Church and C. S. ReVelle, "The maximal covering location problem," *Papers Regional Sci. Assoc.*, vol. 32, pp. 101–118, 1974.

[15] D. J. Eaton, M. S. Daskin, D. Simmons, B. Bulloch, and G. Jansma, "Determining emergency medical deployment in Austin, Texas," *Interfaces*, vol. 15, no. 1, pp. 96–108, Jan./Feb. 1985.

[16] C. R. Toregas, R. Swain, C. S. ReVelle, and L. Bergman, "The location of emergency service facilities," *Oper. Res.*, vol. 19, no. 6, pp. 1363–1373, Oct. 1971.

[17] K. Hogan and C. S. ReVelle, "Concepts and applications of backup coverage," *Manag. Sci.*, vol. 34, no. 11, pp. 1434–1444, Nov. 1986.

[18] M. Gendreau, G. Laporte, and F. Semet, "Solving an ambulance location model by Tabu search," *Location Sci.*, vol. 5, no. 2, pp. 75–88, Aug. 1997.

[19] D. A. Schilling, D. J. Elzinga, J. Cohon, R. L. Church, and C. S. ReVelle, "The TEAM/FLEET models for simultaneous facility and equipment siting," *Transp. Sci.*, vol. 13, no. 2, pp. 163–175, May 1979.

[20] M. S. Daskin, "A maximum expected location model: Formulation, properties and heuristic solution," *Transp. Sci.*, vol. 17, no. 1, pp. 48–70, Feb. 1983.

[21] C. ReVelle and K. Hogan, "The maximum availability location problem," *Transp. Sci.*, vol. 23, no. 3, pp. 192–200, Aug. 1989.

[22] T. Andersson and P. Vaerband, "Decision support tools for ambulance dispatch and relocation," *J. Oper. Res. Soc.*, vol. 58, pp. 195–201, Feb. 2007.

[23] M. S. Maxwell, M. Restrepo, S. G. Henderson, and H. Topaloglu, "Approximate dynamic programming for ambulance redeployment," *INFORMS J. Comput.*, vol. 22, no. 2, pp. 266–281, Spring 2010.

[24] T. Andersson, S. Petersson, and P. Värbrand, "Calculating the preparedness for an efficient ambulance health care," in *Proc. IEEE Intell. Transp. Syst. Conf.*, Washington, DC, Oct. 3–6, 2004, pp. 785–790.

[25] G. Nichol, A. S. Detsky, I. G. Stiell, K. O'Rourke, G. Wells, and A. Laupacis, "Effectiveness of emergency medical services for victims of out-of-hospital cardiac arrest: A metaanalysis," *Ann. Emerg. Med.*, vol. 27, no. 6, pp. 700–710, Jun. 1996.

[26] L. L. Culley, D. K. Henwood, J. J. Clark, M. S. Eisenberg, and C. Horton, "Increasing the efficiency of emergency medical services by using criteria based dispatch," *Ann. Emerg. Med.*, vol. 24, no. 5, pp. 867–872, Nov. 1994.

[27] J. J. Clawson, R. L. Martin, and S. A. Hauert, "Protocols vs. guidelines. Choosing a medical-dispatch program," *Emerg. Med. Serv.*, vol. 23, no. 10, pp. 52–60, Oct. 1994.

[28] J. Nicholl, P. Coleman, G. Parry, J. Turner, and S. Dixon, "Emergency priority dispatch systems—A new era in the provision of ambulance services in the U.K.," *Prehosp. Immediate Care*, vol. 3, pp. 71–75, 1999.

[29] N. Thomson, "Emergency medical services in Zimbabwe," *Resuscitation*, vol. 65, no. 1, pp. 15–19, Apr. 2005.

[30] D. E. Persse, C. B. Key, R. N. Bradley, C. C. Miller, and A. Dhingra, "Cardiac arrest survival as a function of ambulance deployment strategy in a large urban emergency medical services system," *Resuscitation*, vol. 59, no. 1, pp. 97–104, Oct. 2003.

[31] M. S. Eisenberg, B. T. Horwood, R. O. Cummins, R. Reynolds-Haertle, and T. R. Hearne, "Cardiac arrest and resuscitation: A tale of 29 cities," *Ann. Emerg. Med.*, vol. 19, no. 2, pp. 179–186, Feb. 1990.

[32] P. A. Curka, P. E. Pepe, V. F. Ginger, R. C. Sherrard, M. V. Ivy, and B. S. Zachariah, "Emergency medical services priority dispatch," *Ann. Emerg. Med.*, vol. 22, no. 11, pp. 1688–1695, Nov. 1993.

[33] Manual: Emergency Medical Services Administrative Policies and Procedures. Subject: EMS Dispatch Policy, 1996.

[34] H. Adams, "Urgences-santé uses GIS to save lives," *ArcNorth News*, vol. 10, no. 1, p. 13, 2007.

[35] M. Hougham, "London ambulance service computer-aided despatch system," *Int. J. Project Manag.*, vol. 14, no. 2, pp. 103–110, Apr. 1996.

[36] D. W. Spaite, E. A. Criss, T. D. Valenzuela, and J. Guisto, "Emergency medical service systems research: Problems of the past, challenges of the future," *Ann. Emerg. Med.*, vol. 26, no. 2, pp. 146–152, Aug. 1995.

[37] H. K. Rajagopalana, C. Saydamb, and J. Xiao, "A multiperiod set covering location model for dynamic redeployment of ambulances," *Comput. Oper. Res.*, vol. 35, no. 3, pp. 814–826, Mar. 2008.

[38] H. K. Rajagopalana, C. Saydamb, and J. Xiao, "A multiperiod expected covering location model for dynamic redeployment of ambulances," in *Proc. 16th Mini-EURO Conf. 10th Meet. EWGT.*, Poznañ, Poland, Sep. 13–16, 2005, pp. 621–631.

[39] D. J. Lockey, "Prehospital trauma management," *Resuscitation*, vol. 48, no. 1, pp. 5–15, Jan. 2001.

[40] L. M. Beillon, B.-O. Suserud, I. Karlberg, and J. Herlitz, "Does ambulance use differ between geographic areas? A survey of ambulance use in sparsely and densely populated areas," *Am. J. Emerg. Med.*, vol. 27, no. 2, pp. 202–211, Feb. 2009.

[41] A. Khorram-Manesh, K. Lennquist Montán, A. Hedelin, M. Kihlgren, and P. Örtenwall, "Prehospital triage, discrepancy in priority-setting between emergency medical dispatch centre and ambulance crews," *Eur. J. Trauma Emerg. Surg.*, pp. 1–6, May 2010. [Online]. Available: http://www.springerlink.com/content/n260l70380m154t6/export-citation/

[42] A. O'Cathain, E. Webber, J. Nicholl, J. Munro, and E. Knowles, "NHS direct: Constency of triage outcomes," *Emerg. Med. J.*, vol. 20, no. 3, pp. 289–292, May 2003.

[43] P. Toth and D. Vigo, "An overview of vehicle routing problem," in *The Vehicle Routing Problem*. Philadelphia, PA: SIAM, 2002, pp. 1–26.

[44] S. Kim, M. E. Lewis, and C. C. White, III, "State space reduction for nonstationary stochastic shortest path problems with real-time traffic information," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 3, pp. 273–284, Sep. 2005.

[45] I. Chabini and S. Lan, "Adaptations of the A^* algorithm for the computation of fastest paths in deterministic discrete-time dynamic networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 3, no. 1, pp. 60–74, Mar. 2002.

[46] T. L. H. Nguyen, T. H. T. Nguyen, S. Morita, and J. Sakamoto, "Injury and pre-hospital trauma care in Hanoi, Vietnam," *Injury*, vol. 39, no. 9, pp. 1026–1033, Sep. 2008.

[47] M. K. Joshipura, "Trauma care in India: Current scenario," *World J. Surg.*, vol. 32, no. 8, pp. 1613–1617, Aug. 2008.

[48] O. C. Kobusingye, A. A. Hyder, D. Bishai, E. R. Hicks, C. Mock, and M. Joshipura, "Emergency medical systems in low- and middle-income countries: Recommendations for action," *Bull. World Health Org.*, vol. 83, no. 8, pp. 626–631, Aug. 2005.

[49] M. Hauswald and E. Yeoh, "Designing a prehospital system for a developing country: Estimated cost and benefits," *Amer. J. Emerg. Med.*, vol. 15, no. 6, pp. 600–603, Oct. 1997.