Journal of the Operational Research Society (2011) 62, 1888–1897

www.palgrave-journals.com/jors/

The role of preparedness in ambulance dispatching

S Lee*

School of Industrial Engineering, Purdue University, Indiana, USA

Response time in the emergency medical service is an important performance measure and ambulance dispatching is one of the most important factors affecting the response time. The most commonly used dispatching rule is to send the closest available unit to the call site. However, though dispatching the closest unit enables the service to achieve the minimal response time for the current call, the response times for the next incoming calls may increase if the area where the closest ambulance is currently located has a high call rate, that is the area becomes ill-prepared. A dispatching algorithm based on the preparedness concept was recently proposed. Rather than greedily minimizing each current response time, the dispatching algorithm takes account of future calls by a quantitative definition of preparedness. This study investigates the role of preparedness by examining the performance of the preparedness-based dispatching algorithm as well as by evolving the algorithm in several ways in order to magnify the effectiveness of preparedness consideration. As a result of these efforts, it is found that the consideration of preparedness in ambulance dispatching can provide significant benefits in reducing response time but only when appropriately used.

Journal of the Operational Research Society (2011) **62,** 1888–1897. doi:10.1057/jors.2010.145
Published online 3 November 2010

Keywords: ambulance dispatching; preparedness; composite performance indicator; social welfare functions

1. Introduction

When there is an emergency call, an ambulance dispatcher dispatches an ambulance(s) to the patient(s) based on the information gathered such as urgency and medical needs. Response time in the emergency medical service is the time elapsed after an emergency call until an ambulance reaches the patient. The response time has been used as an important performance measure since it directly affects the survival rate in critical emergency situations. For example, after the occurrence of a cardiac arrest, the reduction of every 1 min in response time increases the survival rate by about 10% (BBC News, 2002). To account for the importance of the response time, the fraction of calls whose response time is within a certain time limit is commonly used as the performance indicator. The United State Emergency Medical Services Act provides some standards on the response time: 95% of calls should be responded within 10 min in urban areas while in rural areas within 30 min (Ball and Lin, 1993).

Three types of decisions in the emergency medical service significantly influence the response time: ambulance location, ambulance relocation, and ambulance dispatching. Ambulance location problems have been the most widely studied trying to determine optimal locations for ambulance stations, for example refer to detailed reviews by ReVelle and Hogan (1989), Brotcorne *et al* (2003), Goldberg (2004), and Jia *et al* (2007). The fraction of calls responded within a time limit is often approximated using so-called coverage, assuming a demand node is covered by an ambulance station if the time length between them is within the time limit. Relocation decisions enforce ambulances to move to different locations in order to increase the coverage based on temporal and geographical demand patterns (Kolesar and Walker, 1974; Gendreau *et al*, 2001; Gendreau *et al*, 2006). The relocation decisions need to be made quickly and hence algorithmic efficiency here is an important consideration.

Ambulance dispatching decisions assign appropriate ambulances to the calls and the most common dispatching rule used in practice is to send the closest unit available (Hayes *et al*, 2004; Dean, 2008). This policy is rational since the objective is to minimize the response time. The contributions in ambulance dispatching are sparse, but an argument can be found in the literature against this policy, which was originally made by Carter *et al* (1972) and thereafter supported by Cuninghame-Green and Harries (1988), Repede and Bernardo (1994), and Weintraub *et al* (1999). Though dispatching the closest unit enables to achieve the minimum response time for the current call, the response times for the next incoming calls may increase when the closest ambulance is currently located in an area with a high rate of calls. Dispatching that ambulance significantly reduces the preparedness of the area for the future calls and it would be more desirable

*Correspondence: S Lee, School of Industrial Engineering, Purdue University, 315 N. Grant Street, West Lafayette, Indiana 47907, USA.
E-mail: lee46@purdue.edu

to dispatch a unit located in a farther area with a relatively low rate of calls.

Andersson and Värbrand (2007), for the first time, proposed a dispatching algorithm based on a quantitative definition of preparedness, briefly described as follows and detailed in the next section. A service area is divided into zones and each zone is evaluated in ‘zone preparedness’. A zone is considered more prepared when more ambulances are available in closer distances and call rate is lower. The zone preparedness levels are aggregated into ‘area preparedness’ by taking the minimum zone preparedness level. The area preparedness represents the overall preparedness level of entire service area and is used to guide the dispatching decisions. When there are alternative ambulances for a call, the area preparedness level resulting from dispatching each alternative is computed and the ambulance resulting in the maximum area preparedness is dispatched to the call. This process requires intensive use of computation and communication capabilities and hence advanced information technologies available in these days should be utilized to realize it in practice.

Rather than greedily minimizing each current response time, the dispatching algorithm mentioned above takes account of future events and hence has a potential to reduce the response time. This study investigates the role of preparedness by examining the performance of the preparedness-based dispatching algorithm as well as by evolving the algorithm in several ways in order to magnify the effectiveness of preparedness. There are two modifications made upon the original preparedness-based dispatching algorithm. One is the use of the composite performance indicator incorporating both greediness (used in the conventional algorithm of dispatching the closest unit) and preparedness (used in the preparedness-based algorithm), and another is the utilization of social welfare functions in aggregating zone preparedness levels into area preparedness. As a result of this effort, it is found that the preparedness consideration is useful in reducing the response time but such a benefit can be accomplished only when it is appropriately used.

The fundamental findings of this study will contribute to devising preparedness-based dispatching algorithms for priority dispatching systems where the priority of a call is determined by the urgency of the call. The priority dispatching is increasingly adopted in many countries but it requires more extensive analysis as will be discussed in the concluding section. This paper is organized as follows. Section 2 details the preparedness-based dispatching algorithm and evaluates its performance in various scenarios. The algorithm is reinforced first by devising a composite performance indicator that combines the preparedness with greediness in Section 3, and second by introducing several aggregate functions from social sciences for aggregating zone preparedness levels into area preparedness in Section 4. Finally, Section 5 concludes this work and discusses future work.

2. Preparedness-based dispatching algorithm

Andersson and Värbrand (2007) proposed ambulance dispatching and relocation algorithms based on a quantitative definition of preparedness, and made a simulation study using data from a county of Sweden to test if the algorithms achieve a target performance in different test conditions. This section details the preparedness-based dispatching algorithm, and evaluates it in various scenarios in comparison with the conventional algorithm of dispatching the closest ambulance.

2.1 Algorithm description

The algorithm is based on a quantitative definition of preparedness. A service area is divided into zones Z and the preparedness z_j of zone j (zone preparedness) is mathematically defined as

$$z_j = \frac{1}{\lambda_j} \sum_{i \in A} \frac{\alpha_i}{t_{ij}},$$

where λ_j represents the call rate in zone j, A is a set of available (idle) ambulances, and t_{ij} is the travelling time of ambulance i to zone j. The α_i is the contribution factor of ambulance i and determined such that:

$$t_{1j} \leqslant t_{2j} \leqslant \cdots \leqslant t_{|A|j}$$

$$\alpha_1 > \alpha_2 > \cdots > \alpha_{|A|}.$$

The definition of zone preparedness is rational since a zone is considered more prepared when more ambulances are available in closer distances and the zone has a lower call rate. However, since the parameter space is very large due to the contribution factor α_i, let us reduce the space by using $\alpha_i = 1$ for all i throughout the paper. This simplification is a special case where the contribution factor values are set very close to each other, and preserves the rationality behind the definition of preparedness. In fact, Andersson and Värbrand (2007) used $\alpha_i = 1/2^{i-1}$ as a special form of the contribution factor, and it was confirmed in the preliminary study that both forms of contribution factor do not make any significant difference.

On the basis of the zone preparedness, the dispatching decision is made in five steps as follows:

(i) When a call is received from a zone, a set T of available ambulances are identified whose travelling time to the zone is within a certain threshold δ.
(ii) If there is no ambulance available within the threshold, that is $T = \emptyset$, the closest ambulance (beyond the threshold) is dispatched. Otherwise proceed through steps (iii), (iv), and (v).

(iii) For each ambulance $i \in T$, zone preparedness levels $z_j(i)$ are computed for all the zones $j \in Z$ by setting ambulance i unavailable (resulting from dispatching it).
(iv) The zone preparedness levels resulting from dispatching ambulance $i \in T$ are aggregated into area preparedness $a(i)$ by taking the minimum of zone preparedness levels, that is $a(i) = \min_{j \in Z} z_j(i)$.
(v) The area preparedness $a(i)$ represents the preparedness level of entire service area resulting from dispatching ambulance i, and an ambulance with the maximum area preparedness, that is $\arg\max_{i \in T} a(i)$, is dispatched to the call site.

Since it is not specified in the algorithm what happens if there is no ambulance available in the entire service area, it is assumed in this paper that calls are queued and the ambulance released will serve the call closest to the ambulance. The algorithm is adaptable to priority dispatching systems where the priority of a call is determined by the urgency of the call. The threshold parameter δ in step (i) can be determined based on the priority, for example decrease the threshold for higher priority calls. The priority dispatching is not considered in the current work in order to purely investigate the role of preparedness, but will be discussed as future work in Section 5.

2.2 Performance evaluation

The preparedness-based dispatching algorithm (Preparedness algorithm) described above is evaluated in a discrete event simulator, in comparison with the conventional algorithm of dispatching the closest ambulance (Closest algorithm). The performance is measured in two ways: average response time and cumulative proportion of response time.

2.2.1 Experimental design. The service area is represented by a square grid with 25 vertices as shown in Figure 1, where each vertex (zone) generates calls at a certain rate and ambulances move from vertex to vertex through edges with 1 min of travelling time for every edge. When there is a call, an available ambulance is dispatched to the call site according to a dispatching algorithm adopted. The dispatched ambulance remains unavailable until it reaches the call site and serves the patient. The service time (time taken to serve the patient after reaching the call site) is assumed to follow an exponential distribution with mean = 1 min (Erdoğan *et al*, 2007). The ambulance stays at the call site until dispatched to a different site.

There are two factors that constitute a test condition: (1) call pattern, (2) number of ambulances. Calls are generated in each zone following an exponential distribution (Singer and Donoso, 2008) and four different call patterns are designed as shown in Figure 1: uniform, centred, cornered, and bipartite. A value in the figure represents the call rate of corresponding zones. For example, in the centred pattern, the zone in the centre generates 0.4 calls per min. The values are set such that the entire service area generates 1 call per min. The number of ambulances is in the set $\{2, 4, 6, 8, 10\}$. Resource availability can be defined as the ratio of available resources to workload, and it is controlled here by the available resources (number of ambulances) with the workload (aggregate call rate and size of service area) fixed. Ambulances are initially located at the vertices in the first row (left to right) and then in the second row (left to right) as needed. As a result, 20 test conditions (4 call patterns $\times$ 5 numbers of ambulances) are established.

For each test condition, two dispatching algorithms (Closest and Preparedness) are applied. Since the Preparedness algorithm has threshold parameter δ, four threshold values ($\delta = 1, 3, 5, \infty$) are used. One thousand calls in total are generated for each simulation run and 100 replications are made for each scenario.

2.2.2 Results. Figure 2 shows average response time and Figure 3 shows average cumulative proportion of response time (fraction of calls whose response time is

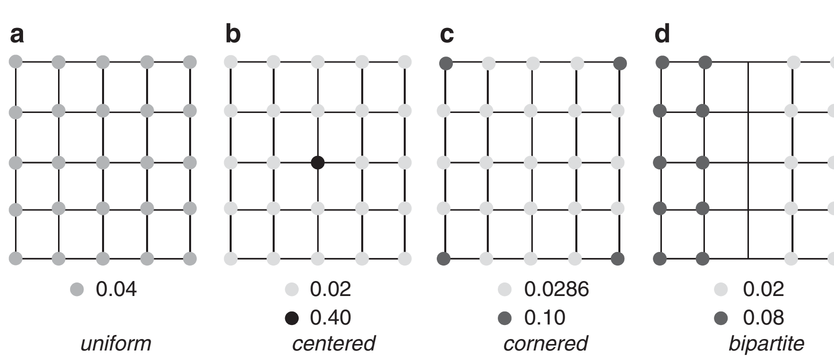

Figure 1 Call patterns over the service area.

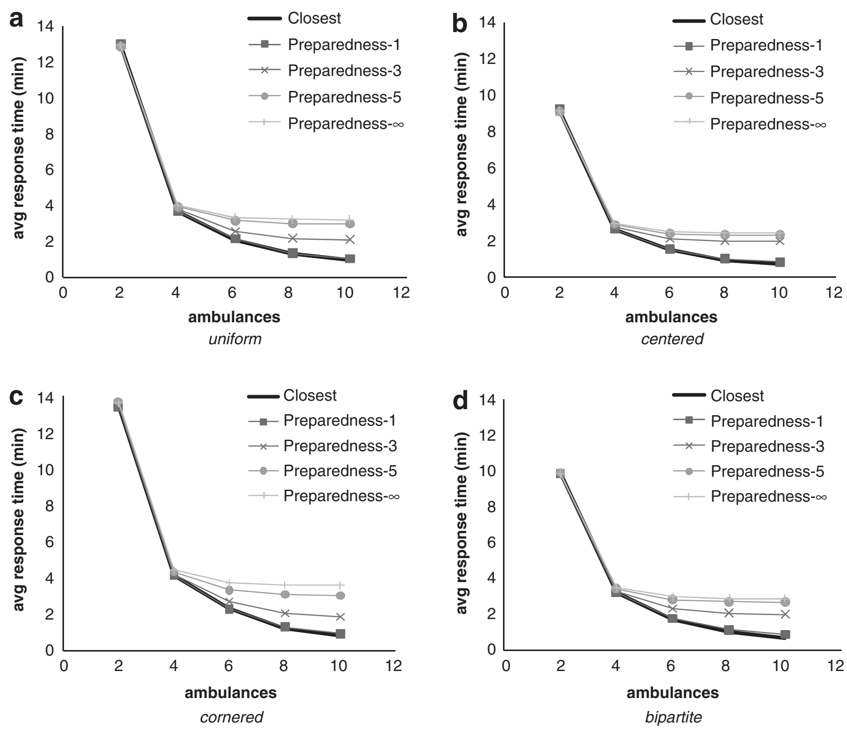

Figure 2 Average response time of the preparedness algorithm.

within x min), in four different call patterns. The Preparedness algorithm performs significantly worse than the Closest algorithm in both measures in all different test conditions. The relative performance degrades with the increase of threshold δ and with the increase of the number of ambulances. Note that the Closest algorithm is a special case of the Preparedness algorithm with $\delta = 0$, and a larger threshold tends to increase the chance of making dispatching decisions based on the preparedness. Similarly, a larger number of ambulances also increases that chance. Therefore, increased consideration of preparedness tends to negatively affect the performance in the current form of the Preparedness algorithm. The concept of preparedness is sound and the algorithm needs to be reinforced to get its potential benefits.

3. Composite performance indicator

In the Preparedness algorithm, preparedness is used as a performance indicator and an ambulance that results in the maximum preparedness is dispatched to the call site. In this section the performance indicator is reinforced into a composite one and its resultant performance is examined.

3.1 Composite algorithm description

The Closest algorithm greedily minimizes each current response time while the Preparedness algorithm tries to increase the preparedness for uncertain future calls. However, since it is more desirable to consider both current (greediness) and future (preparedness) calls, a better strategy would be to use a composite performance indicator that combines both greediness and preparedness. The composite indicator adopted here is $a(i)/t_{ic}$ where t_{ic} is the travelling time of ambulance i to call site c. The Closest algorithm makes decisions only based on t_{ic} and the Preparedness algorithm only on $a(i)$. The modified algorithm, called the Composite algorithm, is the same as the Preparedness algorithm except step (v) that is changed into:

(v) The area preparedness $a(i)$ represents the preparedness level of the entire service area resulting from dispatching ambulance i, and an ambulance that maximizes the composite performance indicator $a(i)/t_{ic}$, that is arg $\max_{i \in T} a(i)/t_{ic}$, is dispatched to the call site c.

3.2 Performance evaluation

The Composite algorithm is evaluated in the same test conditions as described in Section 2.2.1. The composite performance indicator significantly improves the performance of the Preparedness algorithm, close to the one of the Closest algorithm. In order to increase the resolution of comparisons, the performance graphs are presented in different units as shown in Figures 4 and 5. Average response time of the Composite algorithm is represented relatively to the one of the Closest algorithm by *increase in*

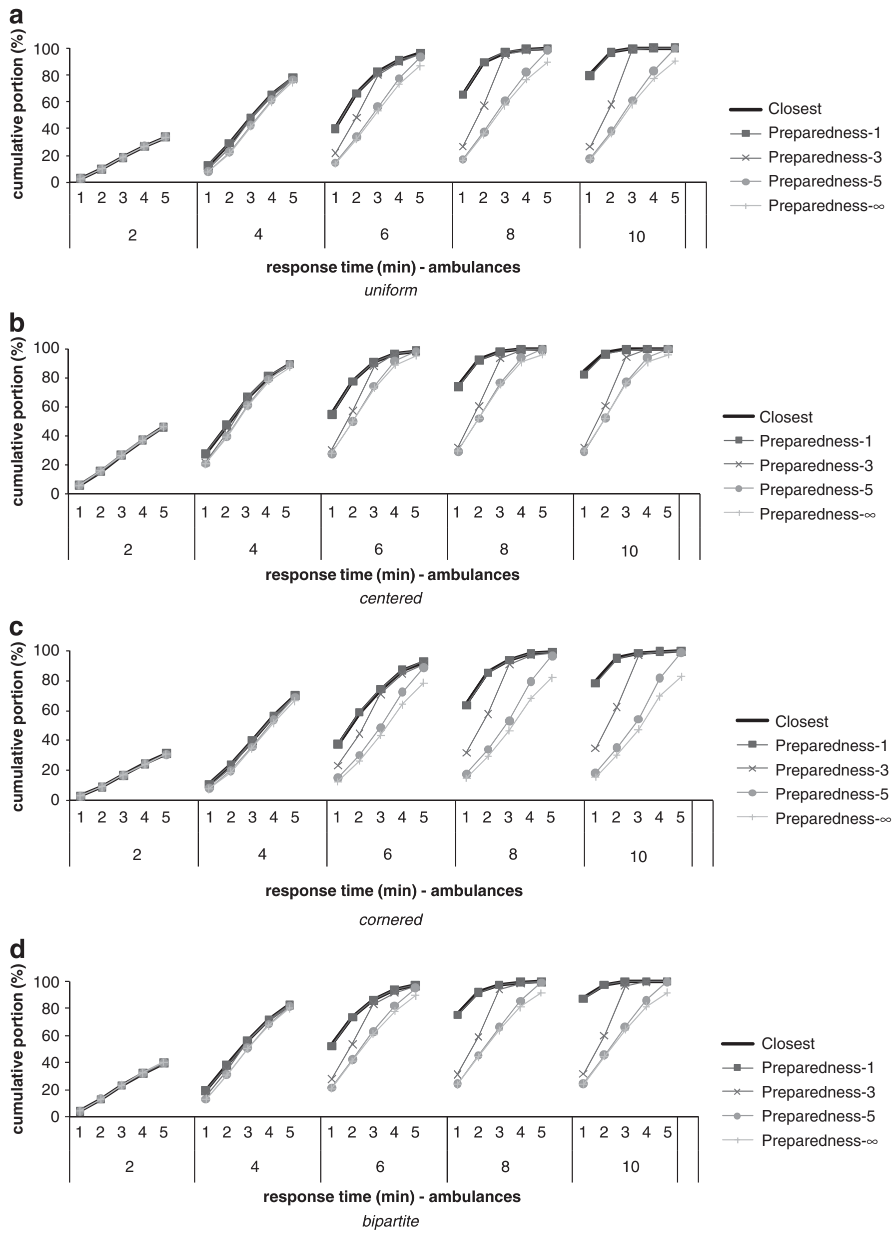

Figure 3 Average cumulative proportion of response time of the preparedness algorithm.

response time = (response time with Composite−response time with Closest)/response time with Closest. Positive values here indicate performance degradation whereas negative values indicate performance improvement. The average cumulative proportion is also represented relatively to that of the Closest algorithm by *increase in cumulative proportion = cumulative proportion with Composite−cumulative proportion with Closest*. Here, positive values indicate performance improvement whereas negative values indicate performance degradation.

Overall, the composite performance indicator improves the performance in both measures beyond the Closest algorithm and it becomes more apparent with the increase of the number of ambulances, now demonstrating the positive impact of preparedness consideration. However, the Composite algorithm is noticeably inferior to the Closest algorithm when there are no more than six ambulances in centred call pattern. This indicates that the Composite algorithm needs to be more robust to different operating environments.

4. Alternative aggregate functions

In the Preparedness algorithm, zone preparedness levels are aggregated into area preparedness by taking the minimum

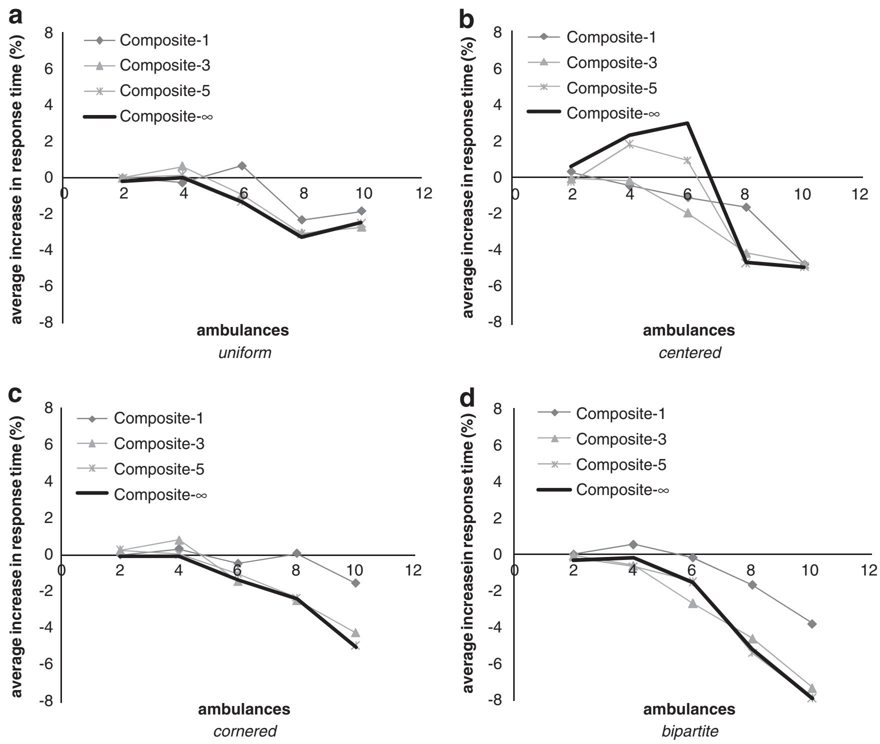

Figure 4 Average response time of the composite algorithm.

zone preparedness level. Apparently there can be different ways of the aggregation, and several aggregate functions are surveyed and evaluated in this section.

4.1 Social welfare functions

When aggregating zone preparedness levels into area preparedness, the Preparedness algorithm takes the minimum zone preparedness level and the aggregate function adopted here is named the Misery function. However, the Misery function only focuses on the minimal zone preparedness level without any consideration of all other zones. One can find different types of aggregate functions in social sciences that can be used to quantify the area preparedness. They are called social welfare functions used to compare welfare between space and time. Average is still the most widely used welfare function despite its well-known shortcomings. Two well-known welfare functions are introduced here that jointly consider average and inequality to arrive at better measures of welfare than average alone.

The Sen welfare function (Sen, 1982), $\bar{P}(1-I_G)$, has a simple form of weighting the average income $\bar{P}$ by Gini index I_G. The Gini index is one of the most commonly used indicators of income inequality. It is derived from a Lorenz curve, which plots the cumulative proportion of income earned by the people ranked from bottom to top as shown in Figure 6. In perfect equality the Lorenz curve follows 45° line. As the degree of inequality increases, the area between the curve and 45° line becomes larger. If the area between the curve and 45° line is A and the area below the curve is B, then the Gini index is computed as A/(A + B). Dagum welfare function (Dagum, 1990), $\bar{P}(1-I_G)/(1+I_G)$, imposes more penalty for inequality on the Sen welfare function through the denominator.

These two social welfare functions can be used for the aggregation with a potential to better represent the area preparedness, since they consider both average and inequality simultaneously. Average function can also be used but it may not perform well because of its indifference to the much less prepared zones. Four aggregate functions have been identified so far and the Preparedness algorithm can use one of the functions with step (iv) changed accordingly. Each function gives rise to a dispatching algorithm whose name is according to the function name as listed below. $\overline{z(i)}$ and $I_G(i)$ are, respectively, the average and Gini inequality of zone preparedness levels resulting from dispatching ambulance *i*. The $I_G(i)$ here is computed from the Lorenz curve that plots the cumulative proportion of zone preparedness levels ranked from bottom to top. Note that all the algorithms defined here use the composite performance indicator devised in the previous section, hence the Misery algorithm is the same as the Composite algorithm.

- Misery algorithm: $a(i) = \min_{j \in Z} z_j(i)$
- Sen algorithm: $a(i) = \overline{z(i)}(1 - I_G(i))$

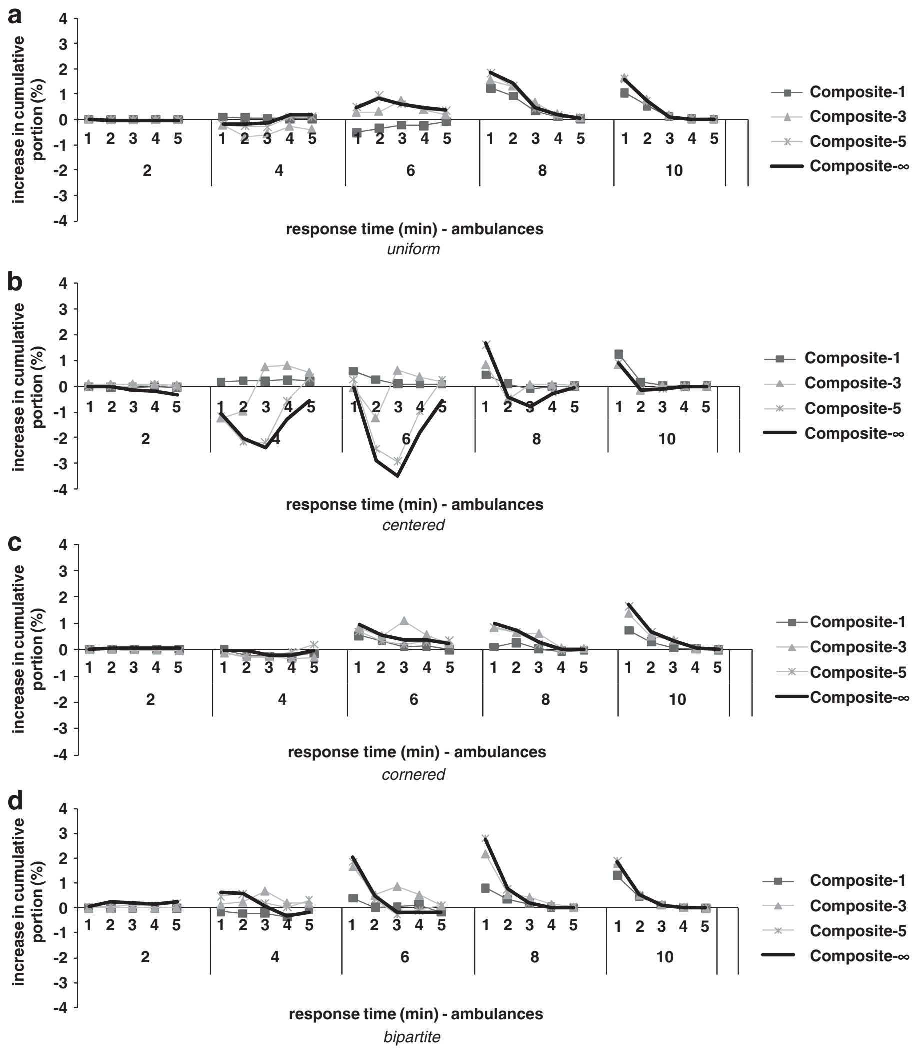

Figure 5 Average cumulative proportion of response time of the composite algorithm.

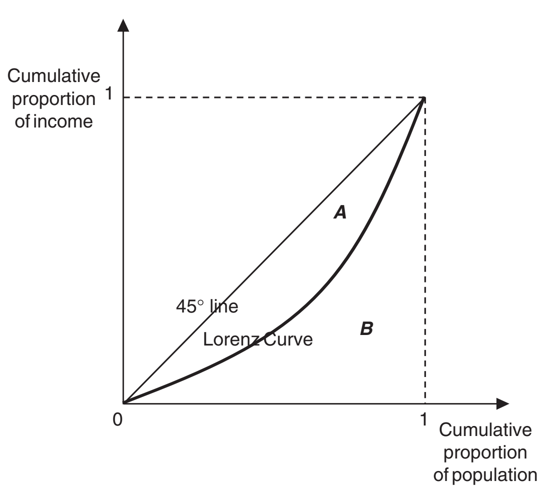

Figure 6 Lorenz curve.

- Dagum algorithm: $a(i) = \overline{z(i)}(1 - I_G(i))/(1 + I_G(i))$
- Average algorithm: $a(i) = \overline{z(i)}$

4.2 Performance evaluation

The four aggregate functions mentioned above are evaluated in the same test conditions as before with threshold (δ) set to ∞ to maximally incorporate the impact of preparedness consideration. The performance graphs are shown in Figures 7 and 8, presented in units the same as the previous section. The Dagum function results in the most preferable performance in uniform call pattern and both Sen and Dagum functions are the most desirable in centred call pattern. The Misery function performs the best in cornered and bipartite call patterns. The Average function, in general, gives inferior performance to others in particular even worse than Closest algorithm.

Overall, the Misery, Sen, and Dagum functions are all plausible candidates capable of representing the

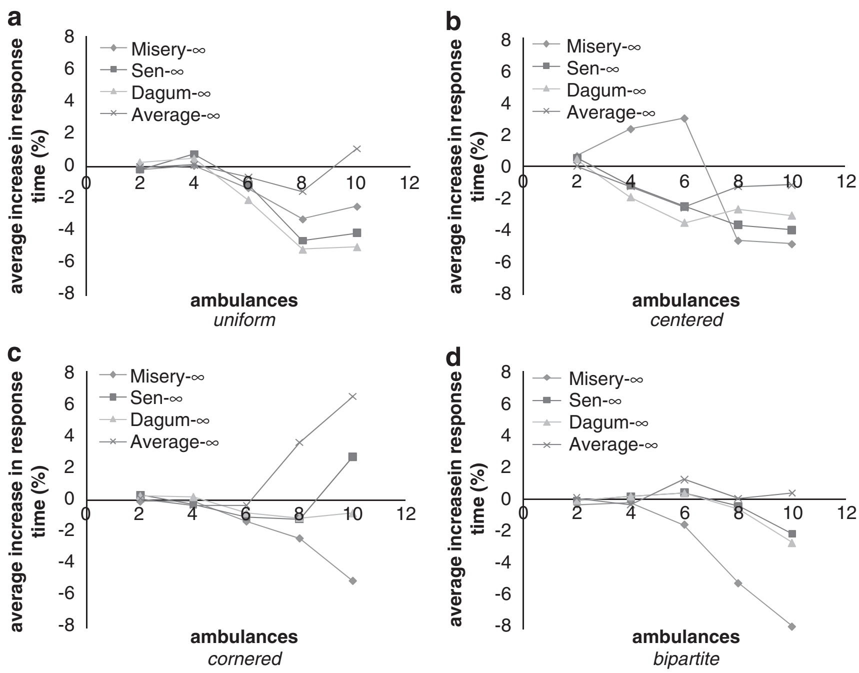

Figure 7 Average response time from different aggregate functions.

preparedness level of entire service area. However, the right function among them should be chosen appropriate to the operating environment factoring in, such as, call pattern and number of ambulances. The right choice of the function gives up to 7.9% decrease (Misery in bipartite) in response time and 3.0% (Dagum in uniform) increase in cumulative proportion, while the wrong choice results in up to 6.4% increase (Average in cornered) in response time and 3.5% decrease (Misery in centred) in cumulative proportion, on the basis of the performance of the conventional Closest algorithm. Also, depending on the choice of the aggregate function the performance can differ by 11.5% (Misery *versus* Average in cornered) in response time and 4.4% (Dagum *versus* Misery in centred) in cumulative proportion.

5. Conclusions and future work

The consideration of preparedness in ambulance dispatching can provide significant benefits in reducing response time beyond the conventional rule of dispatching the closest ambulance, when preparedness is appropriately used. First, preparedness should be used combined with greediness used in the conventional rule, to consider both current (greediness) and future (preparedness) calls. While dispatching decisions only based on preparedness seriously degrade the performance, the composite performance indicator combining preparedness with greediness leads to even better performance than the conventional rule. Second, the function that aggregates preparedness distribution of service area should be capable of quantifying the preparedness level of the entire service area. The Misery, Sen, and Dagum welfare functions are all plausible candidates with such a capability, exhibiting desirable performance in various operating environments. Average function incorporates only partial information of preparedness distribution of service area, leading to the performance even worse than conventional rule.

This research studies the role of preparedness consideration, and identifies algorithmic and environmental factors influencing the dispatching performance, using a simple yet general experimental framework. In order to make the findings of this research transferable to real world environments that are much more complex and dynamic, further investigations need to be made to guide the selection of right aggregate function (as well as threshold). The performance of a dispatching algorithm exhibits nonlinear behaviour and hence an appropriate algorithm needs to be chosen according to the operating environment factoring in, such as, call pattern, available resources, and size of service area. For example, if the size of service area becomes larger, the misery function would turn incapable of representing the preparedness of entire area.

One way is to use an offline approach that finds the best algorithm based on a simulation-based study, which is appropriate for static and predictable environments. However, if the environment is dynamic and unpredictable, it would be more plausible to utilize a learning technique such as Q-learning (Watkins, 1989; Sutton *et al*, 1992;

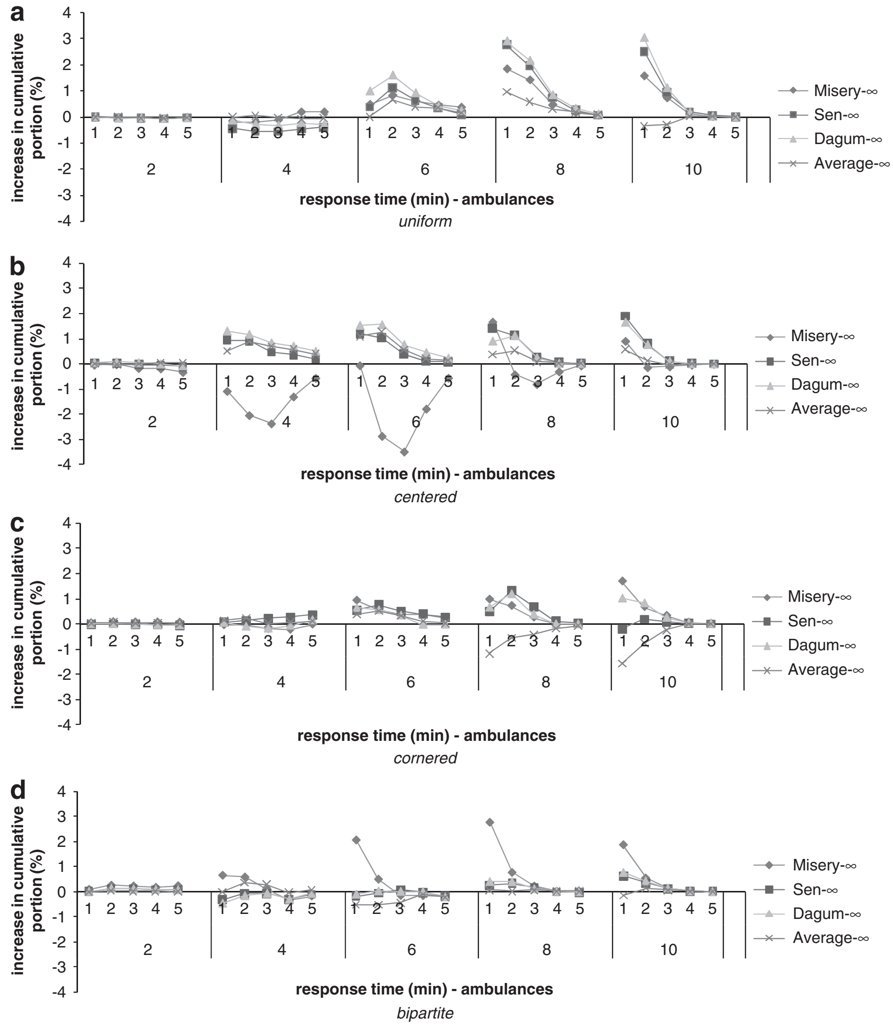

Figure 8 Average cumulative proportion of response time from different aggregate functions.

Barto *et al*, 1995; Kaelbling *et al*, 1996). Q-learning is an online approach that forms optimal policies adaptive to changing environments without explicit knowledge of system models. The alternative aggregate functions are rewarded based on their performance over time and the best function is autonomously found adaptive to the changes of call pattern or number of available ambulances.

Priority dispatching systems, which prioritize ambulance calls in accordance with their degree of urgency, are not considered in this research. However, though non-priority dispatching systems are commonly used in most Asian countries, advanced ambulance service systems in over 20 countries have adopted a priority dispatching system (Fire Services Department, Hong Kong, 2009). Additional issues in the priority systems include the choice of thresholds for different priority classes of calls. Only the ambulances whose travelling time to a call site is within a certain threshold are evaluated in preparedness. The threshold of a priority class will affect the performance of the class itself as well as other classes depending on their thresholds. The complexity of such interactions makes it highly challenging to identify the structure of desirable thresholds.

References

Andersson T and Värbrand P (2007). Decision support tools for ambulance dispatch and relocation. *J Opl Res Soc* **58**: 195–201.

Ball MO and Lin LF (1993). A reliability model applied to emergency service vehicle location. *Opns Res* **41**: 18–36.

Barto G, Bradtke SJ and Singh SP (1995). Learning to act using real-time dynamic programming. *Artif Intell* **72**(1–2): 81–138.

BBC News (2002). Police improve heart attack survival rates, http://news.bbc.co.uk/2/hi/health/2188852.stm, accessed 12 October 2009.

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

Carter G, Chaiken J and Ignall E (1972). Response areas for two emergency units. *Opns Res* **20**: 571–594.

Cuninghame-Green R and Harries G (1988). Nearest-neighbour rules for emergency services. *Zeitschrift for Operations Research* **32**: 299–306.

Dagum C (1990). On the relationship between income inequality measures and social welfare functions. *J Econometrics* **43**(1–2): 91–102.

Dean SF (2008). Why the closest ambulance cannot be dispatched in an urban emergency medical services system. *Prehospital Disaster Medicine* **23**: 161–165.

Erdoğan G, Erkut E and Ingolfsson A (2007). Ambulance location for maximum survival. *Nav Res Log* **55**(1): 42–58.

Fire Services Department, Hong Kong (2009). Ambulance services: Medical priority dispatch system, http://www.hkfsd.gov.hk/home/eng/source/MPDS_Consultation_document_eng.pdf, accessed 12 October 2009.

Gendreau M, Laporte G and Semet F (2001). A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel Comput* **27**: 1641–1653.

Gendreau M, Laporte G and Semet F (2006). The maximal expected coverage relocation problem for emergency vehicles. *J Opl Res Soc* **57**: 22–28.

Goldberg JB (2004). Operations research models for the deployment of emergency services vehicles. *EMS Mngt J* **1**: 20–39.

Hayes J, Moore A, Benwell G and Wong B (2004). Ambulance dispatch complexity and dispatcher decision strategies: Implications for interface design. *Lect Notes Comput Sc* **3101**: 589–593.

Jia H, Ordonez F and Dessouky M (2007). A modeling framework for facility location of medical services for large-scale emergencies. *IIE Trans* **39**: 41–55.

Kaelbling LP, Littman ML and Moore AW (1996). Reinforcement learning: A survey. *J Artif Intell Res* **4**: 237–285.

Kolesar P and Walker W (1974). An algorithm for the dynamic relocation of fire companies. *Opns Res* **22**: 249–274.

Repede J and Bernardo J (1994). Developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky. *Eur J Opl Res* **75**: 567–581.

ReVelle CS and Hogan K (1989). The maximum availability location problem. *Transport Sci* **23**: 192–200.

Sen K (1982). *Choice, Welfare, and Measurement*. Basil Blackwell: Oxford.

Singer M and Donoso P (2008). Assessing an ambulance service with queuing theory. *Comput Opns Res* **35**: 2549–2560.

Sutton RS, Barto AG and Williams RJ (1992). Reinforcement learning is direct adaptive optimal control. *IEEE Contr Syst* **12**(2): 19–22.

Watkins C (1989). *Learning from delayed rewards*. PhD Thesis, King's College: Cambridge University, UK.

Weintraub A, Aboud J, Fernandez C, Laporte G and Ramirez E (1999). An emergency vehicle dispatching system for an electric utility in Chile. *J Opl Res Soc* **50**: 690–696.

Received October 2009;
accepted August 2010 after four revisions