Journal of the Operational Research Society (2010) 61, 543–550 © 2010 Operational Research Society Ltd. All rights reserved. 0160-5682/10

www.palgrave-journals.com/jors/

Scheduling ambulance crews for maximum coverage

G Erdoğan[1], E Erkut[1], A Ingolfsson[2] and G Laporte[3]*

[1]Ozyegin University, İstanbul, Turkey; [2]University of Alberta, Alberta, Canada; and
[3]HEC Montréal, Montréal Canada

This paper addresses the problem of scheduling ambulance crews in order to maximize the coverage throughout a planning horizon. The problem includes the subproblem of locating ambulances to maximize expected coverage with probabilistic response times, for which a tabu search algorithm is developed. The proposed tabu search algorithm is empirically shown to outperform previous approaches for this subproblem. Two integer programming models that use the output of the tabu search algorithm are constructed for the main problem. Computational experiments with real data are conducted. A comparison of the results of the models is presented.

Journal of the Operational Research Society (2010) **61,** 543–550. doi:10.1057/jors.2008.163
Published online 4 February 2009

Keywords: ambulance location; shift scheduling; tabu search; integer programming

1. Introduction

The effectiveness of emergency medical services (EMS) is a crucial ingredient of an efficient healthcare system. The quality of service for EMS systems is measured according to multiple criteria, including average response time, the type of care that EMS staff are trained to provide, and the equipment to which they have access. The most commonly used indicator of quality of service is the fraction of calls whose response time is within a time standard, typically 8–10 minutes. In planning models, this quantity is often approximated using the concept of *coverage*, where a demand node is assumed to be covered by an ambulance station if the average response time is within a preset limit. Many studies exist on improving the quality of service of EMS systems. We refer the reader to Goldberg (2004) for a recent review. Although the basic assumptions of such studies vary, based on the perspective of the modeller and the EMS system at hand, one common assumption is the static availability of the service resources. In other words, once an ambulance is introduced into the system, it is assumed to be active at any given time. However, this is usually not the case in real-world applications because of the human element involved. As prescribed by laws and legislations governing the working conditions of EMS staff, there are limits on the amount and the periods of time an ambulance crew can work in a day. Consequently, there usually exists a limited number of working hour patterns called *shifts* resulting in time-varying service resources. Scheduling the working hours of EMS staff based on these shifts to maximize coverage is a problem that arises periodically. Even when the scheduling decisions are made, the subproblem of locating ambulances based on the varying number of calls still remains unsolved. When implementing results based on existing ambulance location methods, it is crucial to account for the dynamic availability of resources over time.

The purpose of this paper is to develop a solution method for the combined problem of scheduling the working hours of ambulance crews for a given planning horizon and allocating the ambulances to stations distributed throughout a geographical region. The objective is to maximize *expected coverage*, taking into account the probabilistic nature of the problem. The core decision is to allocate ambulance crews to shifts, subject to the maximum number of work hours that can be afforded by the decision makers. The output of the crew-shift assignment is the number of ambulances available for every time interval within the planning horizon. The number of ambulances available at a given time interval should respect a lower limit based on the average number of calls arriving within that time interval. Locating these ambulances to stations in order to maximize expected coverage for the time interval while not exceeding the capacity limits of the stations is also a part of the problem.

The complexity resulting from the time element of the problem can be handled by discretizing it, that is, by dividing the planning horizon into equal-length time intervals. In most applications a planning horizon of one week is appropriate for two reasons: (1) the number of emergency calls received behaves in a cyclic manner with a one-week period; (2) shifts are usually planned so that staffing is constant from week to week. The length of the time intervals is typically considered to be 1 hour. A solution method for the shift scheduling problem should be able to assess the result of allocating a given number of ambulances for a given hour of the

*Correspondence: *G Laporte, Canada Research Chair in Distribution Management, HEC Montréal, 3000 Chemin de la Côte-Sainte-Catherine, Montréal, Canada H3T 2A7.*
E-mail: gilbert@crt.umontreal.ca

week in the form of expected number of calls covered within the hour. However, this assessment is an ambulance location problem on its own. Combining a weekly shift scheduling problem and an ambulance location problem for every hour into a single model is likely to result in an intractable model. A useful observation is that once the scheduling decisions are made, the ambulance location problems for each hour become independent of each other. This assumes that ambulances can be moved between stations every hour to achieve an optimal configuration. Based on our observations of real systems, where dynamic redeployment of ambulances in order to maintain coverage is increasingly the norm, we believe this is a reasonable assumption. In systems where it is not possible to move ambulances as frequently, our model will provide an upper limit on the expected number of covered calls. In such systems, it might be necessary to modify the schedule that our model generates to reduce the number of moves.

In order to cope with the complexity of the shift scheduling problem, we propose to solve the ambulance location problem for the combinations of the call density of every hour in a week and every possible number of ambulances, and use the results as an input to the shift scheduling problem. To this end, we develop a tabu search algorithm for the hourly ambulance location problem. We then construct two alternative models for the shift scheduling problem, which use the output of the tabu search algorithm and vary by their objective function. The first model aims at maximizing the *aggregate expected coverage*, that is, the ratio of the sum of the expected number of calls covered to the total number of calls. The second model is a lexicographic biobjective model, in which the first objective is to maximize the minimum expected coverage over every hour, and the second objective is to maximize the aggregate expected coverage.

The remainder of the paper is organized as follows. In Section 2, we review the existing models for the ambulance location and shift scheduling problems. In Section 3, we develop a tabu search algorithm to solve the subproblem of allocating ambulances to stations and compare our results with those of the previous studies. We construct two integer programming models for the main problem in Section 4. Computational results for both models are presented in Section 5 and conclusions follow in Section 6.

2. Review of related literature

In this section we review the existing literature on the ambulance location models and the shift scheduling models.

2.1. *Ambulance location models*

Ambulance location problems have received a great deal of interest. We refer the reader to Swersey (1994), Marianov and ReVelle (1995), Brotcorne *et al* (2003), and Jia *et al* (2007) for detailed reviews of the related literature. The paper by Brotcorne *et al* (2003) identifies 18 different models for ambulance location. The level of sophistication of a model can be evaluated on its ability to handle the probabilistic nature of the problem, that is, how *expected coverage* is computed. The models involving expected coverage follow two tracks:

(1) Incorporating the probability that a station may have no ambulances to respond to a call: if the probability of having an idle EMS vehicle at a given station is p, then the expected coverage for a demand point within the coverage time limit is not 1 but p (eg, Daskin, 1983; Saydam and McKnew, 1985; ReVelle and Hogan, 1989).

(2) Incorporating response time uncertainty: if the probability of responding from the closest station to a demand point within the given time limit is q and if the closest station has an ambulance, then the expected coverage for that demand point is q (Daskin, 1987).

In a model that incorporates both EMS vehicle availability and response time uncertainty, the expected coverage for a unit demand would be pq, assuming the two sources of uncertainty are independent. Goldberg and Paz (1991) were the first, to our knowledge, to formulate a mathematical program that addressed both sources of uncertainty. They allowed ambulance busy probabilities to vary between stations and used pairwise exchange heuristics to optimize expected coverage, as evaluated by the *Approximate Hypercube* (AH) model of Larson (1975). Ingolfsson *et al* (2008) made the same assumptions but used a different solution heuristic, one that iterates between solving a nonlinear integer program and the AH model. We will refer to the problem studied by Goldberg and Paz (1991) and Ingolfsson *et al* (2008) as the *Maximum Expected Coverage Location Problem with Probabilistic Response Times and Station Specific Busy Probabilities* (MEXCLP+PR+SSBP).

Parameters

n number of stations

m number of demand nodes

q number of ambulances

d_i the average number of calls originating at demand node i

c_j the maximum number of ambulances that can be located at station j

p_j the probability that an ambulance located at station j is busy

P_{ij} the probability that an ambulance dispatched from station j covers demand node i

$i(j)$ the jth preferred station for demand node i. The preference order is based on the distance between the station and the demand node, with ties broken randomly

Letting z_j be the number of ambulances located at station j, the problem can be defined as:

$$\text{maximize } s(z_1, \ldots, z_n) \tag{1}$$

subject to

$$\sum_{j=1}^{n} z_j \leqslant q \tag{2}$$

$$z_j \in \{0, 1, \ldots, c_j\} \tag{3}$$

where the objective function $s(z_1, \ldots, z_n)$ in (1) is the expected number of calls covered, constraint (2) sets the total number of ambulances to be allocated, and constraints (3) set upper bounds on the number of ambulances allocated to each station.

The function $s(.)$ has no known closed-form expression and is only defined for non-negative integer values of its arguments. It can be evaluated using the AH model. Alternatively, if one assumes that the status (busy or idle) of one ambulance is independent of the status of all other ambulances (an assumption made, eg, in Daskin, 1983; Goldberg and Paz, 1991), then (1) can be expressed as

$$\text{maximize } \sum_{i=1}^{m} d_i \sum_{j=1}^{n} P_{i,i(j)}(1 - p_{i(j)}^{z_{i(j)}}) \prod_{u=1}^{j-1} p_{i(u)}^{z_{i(u)}} \tag{4}$$

For a recent study comparing the performance of several ambulance location models including MEXCLP+PR+SSBP, we refer the reader to Erkut *et al* (2009).

2.2. *Shift scheduling models*

As underlined in the survey by Goldberg (2004), shift scheduling for ambulances has received almost no attention in the research literature. We refer the reader to Ernst *et al* (2004) for a general review on staff scheduling and rostering. Typically, such models decouple performance evaluation from scheduling, by assuming a set of staffing requirements for each period that will guarantee the quality of service. Two notable exceptions are Thompson (1997) and Koole and van der Sluis (2003), both of whom maximize an aggregate quality of service measure based on an $M/M/s$ queueing model. In contrast, the quality of service measure that we maximize is based on the hypercube queueing model, where the 'servers' are spatially distributed and closed-form expressions are not available.

Like most of the shift scheduling literature, we use a steady-state approximation to evaluate performance in each period. Green *et al* (2001) have investigated such approximations when the system can be modelled as an $M/M/s$ queue and found that although they are often adequate, they are unreliable in certain situations, such as when average service times are relatively long. Ambulances typically take about an hour to handle a call, suggesting that it may be worthwhile to investigate models that incorporate transient effects, but we leave this for future research.

3. Static allocation of ambulances to stations

In this section we present a tabu search algorithm (Glover and Laguna, 1997) to solve the MEXCLP+PR+SSBP. We use the version of the AH model developed by Budge *et al* (2008), which allows for the possibility of multiple ambulances per station to directly compute the expected coverage, $s(.)$ for a given solution, instead of using the approximation in (4). The solution is encoded in a vector z_i, as the model given in the previous section. At every iteration, we consider moving a single ambulance from one station to another.

Parameters

κ number of iterations since the last update of the best solution value

η the maximum number of iterations without updating the best solution

θ number of iterations for which a vertex stays in the tabu list

ζ the maximum number of ambulances, that is, $\zeta = \sum_{j=1}^{n} c_j$

Step 1 (*Initialization*): Construct a vector $(a_1, \ldots, a_\zeta)$ with binary components, corresponding to an actual physical capacity for ambulance storage. The storage space of a station j is represented by entries in the range $[s_j, t_j]$, where $s_1 = 1$, $s_j = \sum_{i=1}^{j-1} c_i + 1$ for $j > 1$, and $t_j = \sum_{i=1}^{j} c_i$. Set the first q components of the vector to be equal to 1, that is, $a_i = 1$, $\forall i \in \{1, \ldots, q\}$. Set the rest of the components to be equal to 0, that is, $a_i = 0$, $\forall i \in \{q+1, \ldots, \zeta\}$. For $i = 1, \ldots, \zeta - 1$, swap the value of the ith component with the value of the kth component, where k is a randomly selected integer from the interval $[i, \zeta]$. Determine the number of ambulances allocated at station j as $z_j = \sum_{i=s_j}^{t_j} a_i$. Evaluate the solution, and record it as the best solution found. Set $\kappa = 1$.

Step 2 (*Termination check*): If $\kappa = \eta$, stop.

Step 3 (*Local search*): For every station i with $z_i > 0$ and $j \neq i$ with $z_j < c_j$, evaluate the allocation resulting after moving an ambulance from i to j, and record the best new allocation as well as the best new allocation for which the station that lost an ambulance is not in the tabu list. If the best new allocation has a higher expected coverage value than the best solution found so far, set the current solution to be the new solution, update the best solution found, and set $\kappa = 1$. Else, set the current solution to be equal to the best new allocation for which the station that lost an ambulance is not in the tabu list. Add the station that received an ambulance to the tabu list.

Step 4 (*Tabu list update*): Increase the tabu tenure of each vertex in the tabu list by one. Remove from the tabu list the vertices having a tabu tenure greater than or equal to θ. Increment κ by 1. Go to Step 2.

Table 1 Per cent improvement of expected coverage for the tabu search algorithm

q	*System-wide busy probability*					
	0.1 (%)	*0.2 (%)*	*0.3 (%)*	*0.4 (%)*	*0.5 (%)*	*0.6 (%)*
5	0.00	0.21	1.40	2.10	3.47	0.00
6	0.00	0.03	0.90	1.62	1.34	1.68
7	0.00	1.60	0.21	2.01	1.06	1.30
8	0.02	0.25	2.96	0.14	0.42	1.18
9	0.37	0.06	1.17	2.51	1.01	1.27
10	0.00	0.58	3.03	0.25	3.45	0.93
11	0.00	0.20	0.82	0.63	0.96	0.88
12	0.00	0.08	0.75	0.00	0.89	2.55
13	0.36	0.15	0.26	0.98	0.89	2.06
14	0.00	0.26	0.15	0.38	1.71	2.77
15	0.00	0.34	0.00	0.00	1.06	2.21
16	0.00	0.41	0.35	1.08	0.77	1.26
17	0.00	0.40	0.00	0.37	2.04	2.43
18	0.00	0.00	0.03	1.95	1.20	1.24
19	0.00	0.00	0.00	1.18	1.27	2.75
20	0.00	0.00	0.00	0.73	3.08	2.05
21	0.00	0.00	0.00	0.22	2.10	2.03
22	0.00	0.00	0.00	0.00	2.06	3.31
23	0.00	0.00	0.00	0.00	1.34	2.37
24	0.00	0.00	0.00	0.07	0.66	2.06
25	0.00	0.00	0.00	0.20	0.36	1.73
Average	0.04	0.22	0.57	0.78	1.48	1.81

We have implemented our tabu search algorithm using C++ on a Linux workstation with a 64-bit AMS Opteron 275 CPU running at 2.4 GHz. We have conducted computational experiments to compare the performance of our algorithm with the algorithm of Ingolfsson *et al* (2008), using real-world data available from http://www.business.ualberta.ca/aingolfsson/data/. The data consist of the average response times and demand intensity for 16 stations and 180 demand nodes from the city of Edmonton, Canada. Our computational experiments involve two dimensions following the example of Erkut *et al* (2009). The first one is the number of ambulances, and the second is the system-wide busy-probability. The demand data is scaled based on the system-wide busy-probability to reflect the corresponding call intensity. For the tabu search algorithm, we have used $\eta = 20$ and $\theta = \lfloor n/2 \rfloor$. We have performed five replications for each experimental design setting, to eliminate the effect of the random initial solution. Notably, the results of the five replications were always the same for all 126 experimental settings except for one. The performance of our algorithm is never worse than that of the algorithm by Ingolfsson *et al* (2008), and is strictly better for 86 out of 126 cases. The average improvement is 0.82%, with the effect becoming more pronounced for higher values of system-wide busy-probability. The average CPU time is 50.07 seconds for our tabu search algorithm, as compared to 106.56 seconds of the algorithm by Ingolfsson *et al* (2008). Table 1 shows the percent improvement in expected coverage when the tabu search algorithm is used instead of the algorithm by Ingolfsson *et al* (2008).

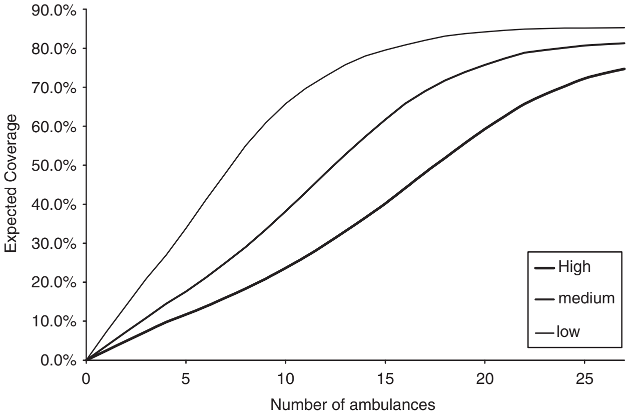

Figure 1 Expected coverage *versus* number of ambulances.

We have extended our experimentation to the rest of the possible number of ambulances and analysed the resulting expected coverage as computed by our algorithm. Three representative results for the cases of low, medium, and high call intensity are depicted in Figure 1. The figure shows that the best bound for the expected coverage increases with the number of ambulances in accordance with an S-curve, which is initially convex and then concave. Initially, the curve is close to linear, but ever so slightly convex. After the inflection point, each additional ambulance adds less additional coverage, because there are fewer and fewer uncovered calls and it becomes increasingly difficult to cover these uncovered calls. The behaviour of the expected coverage as a function of the number of ambulances is similar to, for example, the admission probability as a function of the number of servers in the Erlang B loss model and the no-delay probability as a function of the number of servers in the Erlang C delay model ($M/M/c$ queue) with abandonments. We note that the expected coverage in our model will not necessarily approach 100% as the number of ambulances approaches infinity, because some of the demand locations may be so far from the closest existing station that the probability of coverage will be low no matter how many ambulances are allocated to that station. The asymptotic expected coverage of just over 80% reflects conditions in Edmonton a few years ago, when the city was experiencing rapid growth in population and area, and new stations had yet to be built to accommodate the growth.

4. Weekly scheduling of ambulances

We now turn to the main problem of scheduling ambulance crews. We constructed two integer programming models, their main difference lying in the objective function. Additional notation required to state our models follows.

Notation

δ_{ij} additional number of expected calls covered at hour i by adding the jth ambulance. This value is precomputed as the difference of expected coverage for locating j ambulances and $j-1$ ambulances at the ith hour

h_s the number of working hours for shift s

e_i the average number of calls received during hour i

a_{is} 1 if shift pattern s includes hour i and 0 otherwise

α the average amount of work hours required to serve a call

τ the number of hours in the planning horizon

σ the number of shift patterns

β the benchmark budget, computed as the total amount of work hours required to serve all calls, i.e., $\beta=\alpha\sum_{i=1}^{\tau}e_i$

γ a parameter denoting the amount of budget allocated in terms of the benchmark budget

4.1. Model 1: Maximizing aggregate expected coverage

As stated in the introduction, our first model aims to maximize the aggregate expected coverage, that is, the ratio of the sum of the expected number of calls covered during every hour to the total number of calls. Since we consider coverage to be the primary indicator of quality of service, this model aims to maximize the *performance* of the system. Let x_s be equal to the number of ambulance crews scheduled to work on shift s, and let y_{ij} be equal to 1 if the total number of ambulance crews during hour i is at least j, 0 otherwise. Our first model is then:

(SSP1)

$$\text{maximize} \sum_{i=1}^{\tau}\sum_{j=1}^{\zeta}\delta_{ij}y_{ij}\Big/\sum_{i=1}^{\tau}e_i \tag{5}$$

subject to

$$\sum_{s=1}^{\sigma}a_{is}x_s=\sum_{j=1}^{\zeta}y_{ij} \quad (i\in\{1,\ldots,\tau\}) \tag{6}$$

$$y_{ij}\leqslant y_{i,j-1} \quad (i\in\{1,\ldots,\tau\},\ j\in\{2,\ldots,\zeta\}) \tag{7}$$

$$\sum_{j=1}^{\zeta}y_{ij}\geqslant\lceil\alpha e_i\rceil \quad (i\in\{1,\ldots,\tau\}) \tag{8}$$

$$\sum_{s=1}^{\sigma}h_s x_s\leqslant\lfloor\gamma\beta\rfloor \tag{9}$$

$$x_s\in\mathbb{N} \quad (s\in\{1,\ldots,\sigma\}) \tag{10}$$

$$y_{ij}\in\{0,1\} \quad (i\in\{1,\ldots,\tau\},\ j\in\{1,\ldots,\zeta\}) \tag{11}$$

Constraints (6) set the sum of the number of crews scheduled to shifts that are active during a given hour to be equal to the number of ambulances available in that hour. Constraints (7) state that the jth ambulance can be available only if the $j-1$st is available. Constraints (8) set the lower bound on the ambulances available in a given hour as the number of work hours required to serve all calls in that hour. Although it is conceivable that overall expected coverage could be increased by violating these constraints in certain hours, this would amount to planning to refuse service to some patients, which is unlikely to be acceptable in practice. Note that the constraints (7) for a given hour can be discarded if the constraints (8) put the minimum number of ambulances in that hour above the inflection point of the S-curve. Finally, constraint (9) limits the ambulance crews in terms of maximum work hours that can be afforded. The right-hand side of (9) is stated in a parametric way for ease of experimentation.

4.2. Model 2: Maximin expected coverage, maximum aggregate expected coverage

Although the first model captures the essence of the system at hand, it disregards the concept of *temporal equity*. In order to cover more calls, it could keep the number of ambulances at the bare minimum at hours with low call intensity and place more ambulances at hours with peak call intensity. A remedy to this problem is to maximize the minimum expected coverage over every hour. However, this approach may result in an underutilization of system resources because this alternative objective function does not differentiate between optimal solutions with differing aggregate expected coverage values. Our aim should then be to find the solution with maximum expected coverage *and* aggregate expected coverage. Consequently our second model is lexicographically multiobjective, where the first objective is to maximize the minimum expected coverage over every hour, and the second objective is to maximize the aggregate expected coverage. We write maximize (z_1, z_2) to denote a lexicographic maximization with z_1 being the first objective and z_2 being the second. Let w be equal to the minimum expected coverage over every hour.

(SSP2)

$$\text{maximize}\left(w,\ \sum_{i=1}^{\tau}\sum_{j=1}^{\zeta}\delta_{ij}y_{ij}\Big/\sum_{i=1}^{\tau}e_i\right) \tag{12}$$

subject to

$$w\leqslant\sum_{j=1}^{\zeta}\delta_{ij}y_{ij}/e_i \quad (i\in\{1,\ldots,\tau\}) \tag{13}$$

and (6), (7), (8), (9), (10), and (11).

Solving SSP2 requires solving two integer programming models sequentially, the first of which is simply the model above with the first objective function. Denoting the optimal objective value of the first stage as w^*, the second stage problem is:

$$\text{maximize} \sum_{i=1}^{\tau}\sum_{j=1}^{\zeta}\delta_{ij}y_{ij}\Big/\sum_{i=1}^{\tau}e_i \tag{14}$$

subject to

$$w^*\leqslant\sum_{j=1}^{\zeta}\delta_{ij}y_{ij}/d_i \quad (i\in\{1,\ldots,\tau\}) \tag{15}$$

and (6), (7), (8), (9), (10), and (11).

The first stage maximizes the minimum expected coverage, while the second stage maximizes the aggregate expected coverage subject to the constraint that the minimum expected coverage is greater than or equal to the optimal solution value of the first stage.

We use the output of both models to find the number of ambulances available at each hour. We then allocate these ambulances as determined in the preprocessing phase.

5. Computational results

We used the platform and data described in Section 3 to experiment with the models presented in the previous section. The first part of the experimentation was to run our tabu search algorithm for every hour of the week and every possible number of ambulances in that hour. The computational effort can be reduced by only considering the number of ambulances that satisfy constraint (8). Note that since a new problem is solved for every different hour, response times that depend on the hour can be easily incorporated into this procedure. If the response times are assumed to be the same for every hour, and several hours of the week have the same demand, then one can just do the computations for one of those hours. We have performed a single replication for each instance. This resulted in a total of $7 \times 24 \times \zeta = 7 \times 24 \times 27 = 4536$ runs and required 59.4 CPU hours (2.5 days). Although the computing time is large, every instance of the preprocessing stage is independent of each other and does not require licensed software that allows parallel computation without tedious implementation. On a computing grid consisting of 32 Linux workstations with 64-bit AMD Opteron CPUs, the wall clock time required to complete the preprocessing phase was a little more than two hours.

We emphasize that the models we have presented in the previous section can use the output of every possible solution method for the ambulance location problem. We have used our tabu search algorithm to obtain results that are as realistic as possible. In the case where no more than one CPU can be allocated to the preprocessing stage, one may opt for less sophisticated ambulance location models such as MEXCLP of Daskin (1983) to save computational effort.

The extra piece of information we needed for the second stage was the a_{is} matrix of shift patterns. We used a matrix with 15 shift patterns that correspond to the shifts that are in current use by a Canadian EMS operator in a mid-size Canadian city. The first shift pattern is a 24-hour shift denoting two crews working shifts of either 12 + 12 or 10 + 14 hours using the same ambulance. The next nine shifts are 12-hour patterns with start times at the beginning of every hour from 7 am to 3 pm. The last four shift patterns consist of 10.5-hour shifts, which we approximated as 11-hour shifts, with start times at 6 am, 7 am, 10 am, and 4 pm. Based on the output of the preprocessing stage, we solved both models from the previous section for values of $\gamma \in \{1.5, 2, 2.5, 3, 3.5, 4\}$. Both models involve about 4500 variables and constraints. Using C++ and the callable library of CPLEX 10.1, the average computing times per instance for SSP1 and SSP2 were 8.99 and 27.14 CPU seconds, respectively. Figure 2 compares the aggregate expected coverage achieved by the two models. Both models behave in a similar manner, starting around 40% and converging to 80% at $\gamma = 4$, at which point the system saturates. The average difference is 0.36% and the maximum difference is 1.17% at $\gamma=2$. We conclude that the emphasis on equity does not result in a severe loss in aggregate expected coverage. This lack of a serious conflict between the equity and system performance metrics is certainly good news for EMS planners who must be concerned with both.

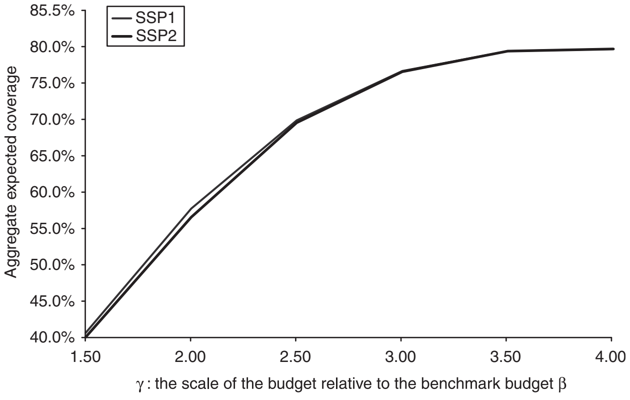

Figure 2 Aggregate expected coverage *versus* γ.

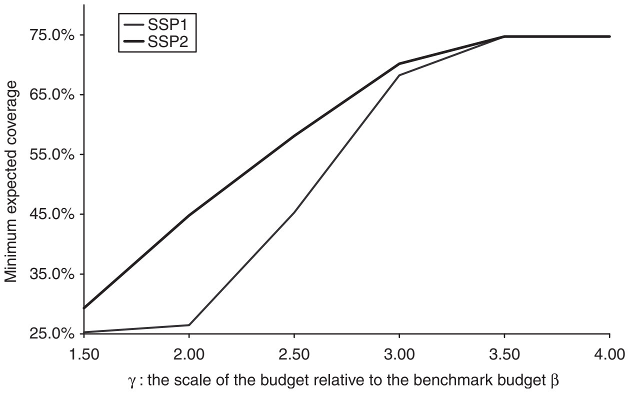

Figure 3 Minimum expected coverage *versus* γ.

Figure 3 compares the minimum expected coverage over all hours of the week for the two models. The difference is more pronounced in this case, with an average of 6.19% and a maximum of 18.34% for $\gamma = 2$. Experiment makes it clear that the lack of equity concerns in SSP1 can result in rather poor solutions from an equity perspective.

We have also analysed the variation of the number of ambulances with respect to call intensity. Figure 4 depicts the comparison of the number of ambulances allocated to each hour of the week by both models, and the pattern of the demand intensity. When the budget is low (γ=1.5), the staffing

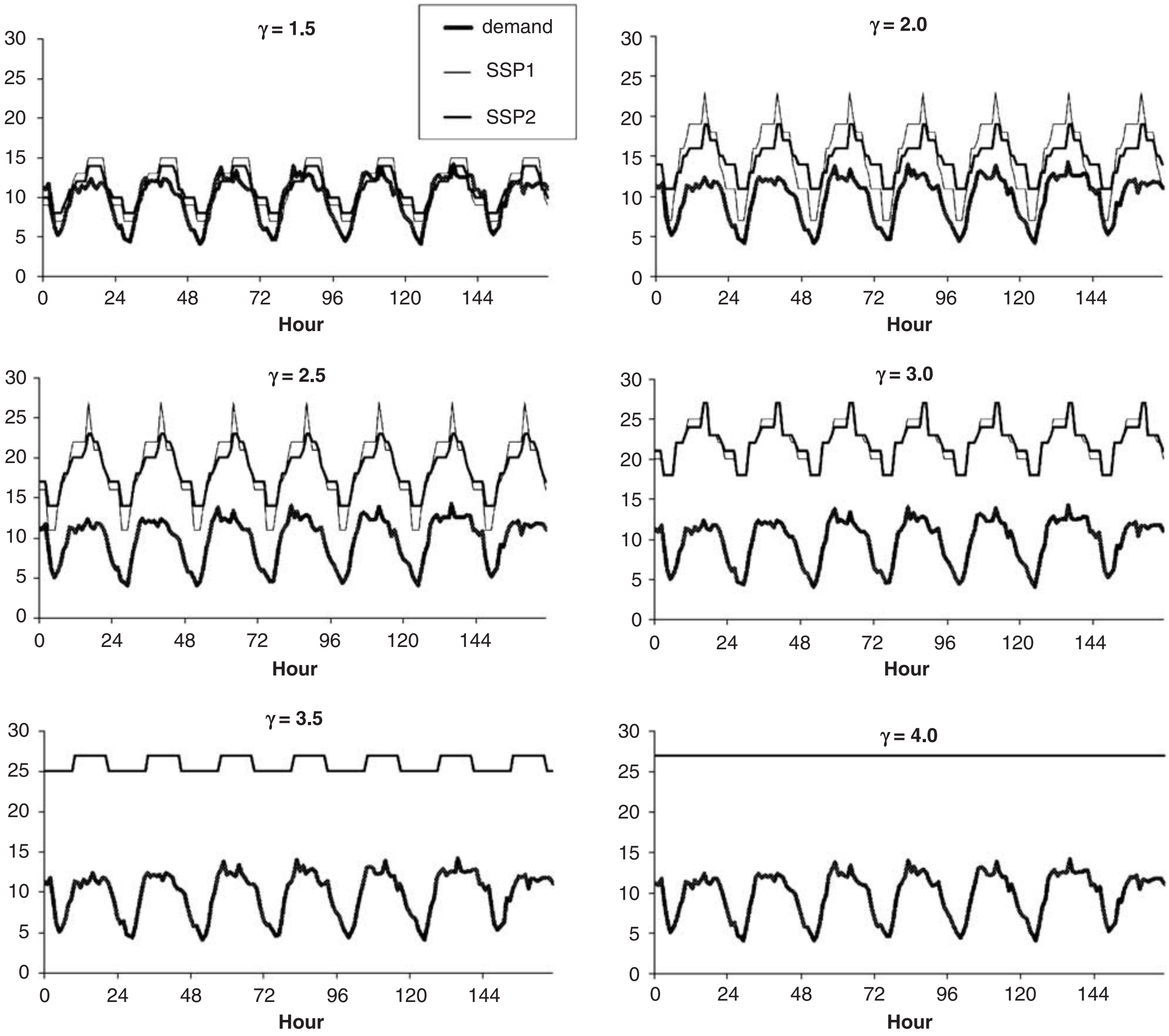

Figure 4 Number of ambulances per hour based on the output of both models. Note that both curves closely follow the demand pattern, with differences in amplitude.

curves are driven primarily by the hourly minimum staffing requirements (8) and are therefore similar for both models. When the budget is large ($\gamma = 3.5$ or 4), the models behave identically. For more realistic intermediate budgets ($\gamma = 2$ or 2.5, corresponding to ambulance utilization of 40 to 50%), the SSP1 model places more emphasis on the peak intensity hours and relatively less emphasis on the low intensity hours. SSP2, on the other hand, is more stable, with a staffing curve that is roughly proportional to the demand intensity.

6. Conclusions

In this study we have analysed the problem of scheduling ambulance crews to shifts in order to maximize coverage. The subproblem of locating the ambulances at stations was solved using a tabu search algorithm, which was empirically shown to outperform the previous approaches in the literature. Two integer programming models were constructed for the problem. Both require the outcome of allocating a given number of ambulances to a given time slot in the planning horizon. The first model emphasizes overall system performance, that is, maximizing the aggregate expected coverage. The second model is a lexicographic biobjective model maximizing temporal equity first, that is, the minimum of hourly expected coverage and the performance second. A computational experiment with real data was conducted. The experiment consists of a parallel preprocessing phase regarding the tabu search algorithm, and running the models on the output of the preprocessing phase. The outputs of the models were graphically analysed. Our results indicate that the second model can handle the maximization of equity with an average of 0.29% and a maximum of 1.44% loss in performance.

Acknowledgements— This work was partially funded by the Canadian Natural Sciences and Engineering Research Council under Grants 39682-05 and 203534-07. This support is gratefully acknowledged. We thank the Edmonton and Calgary EMS departments for providing access to data, and Dan Haight and Matt Stanton of the Centre for Excellence in Operations at the University of Alberta School of Business for their assistance with data preparation. Finally, thanks are due to two referees for their valuable comments.

References

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

Budge S, Ingolfsson A and Erkut E (2008). Approximating vehicle dispatch probabilities for emergency service systems with location-specific service times and multiple units per location. *Opns Res*, forthcoming.

Daskin MS (1983). A maximum expected covering location model: Formulation, properties, and heuristic solution. *Transport Sci* **17**: 48–70.

Daskin MS (1987). Location, dispatching, and routing model for emergency services with stochastic travel times. In: Ghosh A and Rushton G (eds). *Spatial Analysis and Location Allocation Models*. Van Nostrand Reinhold: New York, pp 224–265.

Erkut E, Ingolfsson A, Sim T and Erdoğan G (2009). Computational comparison of five maximal covering models for locating ambulances. *Geogr Anal* **41**: 43–65.

Ernst AT, Jiang H, Krishnamoorthy M and Sier D (2004). Staff scheduling and rostering: A review of applications, methods and models. *Eur J Opl Res* **153**: 3–27.

Glover F and Laguna M (1997). *Tabu Search*. Kluwer Academic Publishers: Boston.

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

Goldberg JB and Paz L (1991). Locating emergency vehicle bases when service time depends on call location. *Transport Sci* **25**: 264–280.

Green LV, Kolesar PJ and Soares J (2001). Improving the SIPP approach for staffing service systems that have cyclic demands. *Opns Res* **49**: 549–564.

Ingolfsson A, Budge S and Erkut E (2008). Optimal ambulance location with random delays and travel times. *Health Care Mngt Sci* **11**: 262–274.

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.

Koole G and van der Sluis E (2003). Optimal shift scheduling with a global service level constraint. *IIE Trans* **35**: 1049–1055.

Larson RC (1975). Approximating the performance of urban emergency service systems. *Opns Res* **23**: 845–868.

Marianov V and ReVelle CS (1995). Siting emergency services. In: Drezner Z (ed). *Facility Location: A Survey of Applications and Methods*. Springer-Verlag: New York, pp 199–222.

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

Saydam C and McKnew M (1985). A separable programming approach to expected coverage: An application to ambulance location. *Decision Sci* **16**: 381–398.

Swersey AJ (1994). The deployment of police, fire, and emergency medical units. In: Barnett A, Pollock SM and Rothkopf MH (eds). *Handbooks in Operations Research and Management Science, Operations Research and the Public Sector*, Vol. 6. North Holland: Amsterdam, pp 151–200.

Thompson GM (1997). Labor staffing and scheduling models for controlling service levels. *Nav Res Log* **44**: 719–740.

Received April 2008;
accepted November 2008 after one revision