A DYNAMIC MODEL AND PARALLEL TABU SEARCH HEURISTIC FOR REAL-TIME AMBULANCE RELOCATION

by

Michel Gendreau [1,3]
Gilbert Laporte [2,3]
Frédéric Semet [3,4,5]

May, 2000

[1] Département d'informatique et de recherche opérationnelle, Université de Montréal, C.P. 6128, succursale Centre-Ville, Montréal H3C 3J7, Canada.
[2] École des Hautes Études Commerciales, 3000 Chemin de la Côte-Sainte-Catherine, Montréal H3T 2A7, Canada.
[3] Centre de recherche sur les transports, Université de Montréal, C.P. 6128, succursale Centre-Ville, Montréal H3C 3J7, Canada.
[4] GRIS, Département d'administration de la santé, Université de Montréal, C.P. 6128, succursale Centre-Ville, Montréal H3C 3J7, Canada.
[5] LAMIH-ROI, Université de Valenciennes, Le Mont-Houy, 59304 Valenciennes Cedex 9, France.

Abstract

This paper considers the redeployment problem for a fleet of ambulances. This problem is encountered in the real-time management of emergency medical services. A dynamic model is proposed and a dynamic ambulance management system is described. This system includes a parallel tabu search heuristic to precompute redeployment scenarios. Simulations based on real-data confirm the efficiency of the proposed approach.

Keywords : emergency vehicles, coverage models, tabu search heuristic, real-time.

Résumé

On considère dans cet article le problème de redéploiement d'une flotte d'ambulances. Ce problème intervient lors de la gestion en temps réel d'un système de véhicules d'urgence. On propose pour ce problème un modèle dynamique et l'on décrit un système de gestion en temps réel de la flotte d'ambulance. Ce système comprend une méthode heuristique parallèle afin de déterminer à l'avance des scénarios de redéploiement. Des simulations produites à partir de données réelles confirment la pertinence de l'approche proposée.

Mots-clefs : véhicules d'urgence, modèles de couverture, heuristique avec recherche tabou, temps réel.

1. Introduction

The main objective of emergency medical services (EMS) is to save lives but the potential of such systems to reduce mortality is related to paramedics training and to the time needed by a paramedic team to arrive on scene. In a real-time context, EMS managers are faced with two main problems: an *allocation problem* and a *redeployment problem*. The allocation problem consists of determining which ambulance must be sent to answer a call. The redeployment problem consists of relocating available ambulances to the potential location sites when calls are received. Basically, the redeployment problem can be viewed as an ambulance location problem. Ambulances are assigned to potential sites to provide adequate coverage. In our case, we consider two types of covering constraint. The absolute covering constraints require that all the demand be satisfied by an ambulance within r_2 minutes, and the relative covering constraints state that a proportion α of the total demand is also within r_1 minutes of an ambulance ($r_2 > r_1$). This type of constraint is in agreement with the United States EMS Act of 1973 in which the standards are $r_1 = 10$ minutes and $\alpha = 0.95$. In Montreal, ambulances are run by "Urgences Santé" which uses $r_1 = 7$ minutes and $\alpha = 0.9$ Desrosiers and Thibault [3]. A standard of $r_2 = 15$ minutes is the objective set by the company for the near future.

The redeployment problem differs from the standard ambulance location problem in several respects. While location problems are usually solved at the *strategic level*, redeployment problems are mainly *operational* and are solved dynamically in real-time, as EMS managers must make almost instantaneous and simultaneous decisions relative to allocation and redeployment. Also, in redeployment problems the current location of vehicles plays a key role, whereas location problems are usually solved from scratch. Finally, a number of additional practical constraints are also present. These include the following: 1) a limited number of ambulances can be positioned at each site; 2) only a limited number of ambulances can be moved when a redeployment occurs; 3) vehicles moved in successive redeployments cannot be always the same; 4) repeated round trips between two location sites must be avoided; 5) long trips between the initial and final location sites must be avoided; 6) an assignment to a call should be avoided near the end of a working shift; 7) at the end of a shift, the ambulance has to be moved closer to the

central service point where the vehicles are based; 8) the breaks of paramedic teams have to be taken into account.

To our knowledge, the redeployment problem has hardly been studied in the past. Most published papers consider the static ambulance location problem only Marianov and ReVelle [11] for a survey. This can be explained mainly by the limitations of past technology which did not allow for real-time solutions of dynamic large-scale problems. The development of new telecommunication and computer technologies now make the redeployment problem tractable since the required data can be obtained and processed in real-time Brotcorne, Farand, Laporte and Semet, [2]. For instance, the positions of vehicles are always available through a Geographic Positioning System (GPS) and can be reported on a computerized map managed by a Geographic Information System (GIS). From an operations research perspective, the advent of new solution methodologies in the fields of parallel computing and metaheuristics opens new research avenues and problems of realistic dimensions can now be solved in real-time. Interesting examples can be found in the area of dynamic vehicle dispatching, for example Gendreau, Guertin, Potvin and Taillard [4]. The use of parallel computing is particularly critical in contexts such as ours where immediate response is required. Not only does parallel computing achieve greater speed, but it also enables the use of more sophisticated algorithmic approaches which would simply be intractable within conventional computing environments.

The remainder of this paper is organized as follows. The model is presented in the next section. In Section 3, a real-time ambulance management system is described. This is followed by computational experiments in Section 4 and by some conclusions in Section 5.

2. A dynamic model

Over the past fifteen years, several coverage models have been put forward for the static ambulance location coverage problem Hogan and ReVelle [8], ReVelle and Hogan [13], Ball and Lin [1], Marianov and ReVelle [10], Gendreau, Laporte and Semet [5], Mandell [9]. In a major departure from previous studies, the model we propose includes a dynamic dimension in the sense that the solution it produces depends on the current state of the system. A comparison of some recent models is provided in Table 1.

Table 1. Comparison of seven ambulance location coverage models

Model	Objective	Coverage constraints	Ambulances
Hogan and ReVelle [8]	Maximize a linear combination of demand covered at least once and at least twice within r_1.	All demand covered within r_2.	One type of vehicle. Total number given.
ReVelle and Hogan [13]	Maximize the total demand covered with an appropriate number of ambulances.	None.	One type of vehicle. Total number given. At most one ambulance per site.
Ball and Lin [1]	Minimize the sum of ambulance fixed costs.	Proportion α of all demand covered within r_1.	One type of vehicle. Decision variable. No constraint on the number of ambulances per site.
Marianov and ReVelle [11]	Maximize the total demand covered with an appropriate site-specific number of ambulances.	None.	One type of vehicle. Total number given. Upper bound on the number of ambulances per site.
Gendreau, Laporte and Semet [5]	Maximize the total demand covered at least twice within r_1.	All demand covered within r_2. Proportion α of all demand covered within r_1.	One type of vehicle. Total number given. Upper bound on the number of ambulances per site.
Mandell [9]	Maximize the total demand covered.	None.	Two types of vehicle. Upper bound on the number of ambulances per site. Ambulance availability is restriced.
Gendreau, Laporte and Semet (this paper)	Dynamically maximize the total demand covered at least twice within r_1, minus a penalty term to reflect the change from the current state of the system.	All demand covered within r_2. Proportion α of all demand covered within r_1.	One type of vehicle. Total number given. Upper bound (≥ 1) on the number of ambulances per site.

The model we propose, called RP^t, solves a version of the redeployment problem at time *t*. It includes the following constraints listed in Section 1: 1) double covering constraints; 2) constraints on the number of ambulances at each site; 3) constraints avoiding to move the same ambulances repeatedly; 4) constraints avoiding round trips; 5) constraints avoiding long trips. The objective is to maximize the backup coverage demand, i.e., the proportion of the demand covered by at least two vehicles within a radius r_1, minus a relocation cost. The first term of the objective function is particularly appropriate in a real-time context since the zone covered by a vehicle assigned to a call may still be covered by another ambulance after the call has been

serviced. The second term ensures that the location plan remains fairly stable throughout the day. The model is truly dynamic since it incorporates new information on the state of the system received at each period *t*.

Formally, the problem is defined on a graph G = (V∪W, E) where V = $\{v_1,...,v_n\}$ and W = $\{v_{n+1},...,v_{n+m}\}$ are two vertex sets representing, respectively, demand points and potential location sites and E = $\{(v_i,v_j)$: $v_i,v_j \in$ V∪W, i < j$\}$ is an edge set. Demand at vertex is equal to λ_i. With each edge is associated a travel time t_{ij}. For $v_i \in$ V and $v_j \in$ W, define:

γ_{ij} as a binary coefficient indicating whether $t_{ij} \leq r_1$ ($\gamma_{ij} = 1$) or not ($\gamma_{ij} = 0$) ;

δ_{ij} as a binary coefficient indicating whether $t_{ij} \leq r_2$ ($\delta_{ij} = 1$) or not ($\delta_{ij} = 0$).

The total number of available ambulances is given and equal to p ($p \leq m$). The maximum number of ambulances that can wait at $v_j \in$ W is p_j . To model constraints iii), iv) and v), define a penalty coefficient $M^t_{j\ell}$ associated with the relocation of ambulance $\ell = 1, \ldots, p$ from its current site at time t to location site $v_j \in$ W. These penalty coefficients reflect ambulance locations at time t as well as information regarding previous redeployments before t. Thus the redeployment of an ambulance that has been frequently moved in the past will be more heavily penalized. In the same vein, round trips and ambulance relocations over a long distance will be discouraged. It is important to realize that the $M^t_{j\ell}$ coefficients are updated at each period t. Finally, α is the proportion of the total demand that must be covered by an ambulance located within r_1 units.

We use the following variables: $y_{j\ell}$ is a binary variable equal to 1 if and only if ambulance ℓ is located at $v_j \in$ W and x_i^k is a binary variable equal to 1 if and only if v_i is covered at least k times. The redeployment problem at time t is then:

$$(RP^t) \qquad \text{maximize } \sum_{i=1}^{n} \lambda_i \, x_i^2 - \sum_{j=1}^{m} \sum_{\ell=1}^{p} M^t_{j\ell} \, y_{j\ell} \qquad (1)$$

$$\text{subject to :} \qquad \sum_{j=1}^{m} \sum_{\ell=1}^{p} \delta_{ij} \, y_{j\ell} \geq 1 \qquad \forall \, v_i \in V \qquad (2)$$

$$\sum_{i=1}^{n} \lambda_i \, x_i^1 \geq \alpha \sum_{i=1}^{n} \lambda_i \qquad (3)$$

$$\sum_{j=1}^{m} \sum_{\ell=1}^{p} \gamma_{ij} \, y_{j\ell} \geq x_i^1 + x_i^2 \qquad \forall \, v_i \in V \qquad (4)$$

$$x_i^2 \leq x_i^1 \qquad \forall \, v_i \in V \qquad (5)$$

$$\sum_{j=1}^{m} y_{j\ell} = 1 \qquad \ell = 1, \ldots, p \qquad (6)$$

$$\sum_{\ell=1}^{p} y_{j\ell} \leq p_j \qquad \forall \, v_j \in W \qquad (7)$$

$$x_i^1 = 0 \text{ or } 1$$
$$x_i^2 = 0 \text{ or } 1 \qquad \forall \, v_i \in V \qquad (8)$$

$$y_{j\ell} = 0 \text{ or } 1 \qquad \forall \, v_j \in W \text{ and } \ell = 1, \ldots, p \qquad (9)$$

In this model, constraints (2) and (3) express the single and the double coverage requirements. The absolute covering constraints (2) state that all demand must be covered within r_2 units. Constraints (3) and (4) express the relative covering requirements. Constraints (3) impose that a proportion α of all demand is covered whereas constraints (4) state that the number of ambulances located within r_1 units should be at least one if $x_i^1 = 1$ or at least two if $x_i^2 = x_i^1 = 1$. Constraints (5) ensure that a demand point cannot be covered twice if it is not covered at least once. Constraints (6) specify that each available ambulance must be assigned to a potential location site. Finally, constraints (7) define an upper bound on the number of vehicles waiting at a location site.

3. A real-time ambulance management system

The system we have developed enables real-time decisions to be made at two levels. The first is the *allocation problem* which consists of determining which ambulance to send to a given call. This is easily achieved through the application of some basic rules and does not require a heavy optimization machinery. At the second level, the ambulance *redeployment problem* must be solved in real-time repeatedly throughout the day.

3.1 The allocation problem

The choice of an ambulance to answer a call depends on the nature of the call. Typically, when a call occurs at the EMS dispatching center, the medical status of the patient is evaluated through a standardized procedure which yields four levels of severity: 1) urgent calls requiring one ambulance; 2) urgent calls requiring several ambulances; 3) less urgent calls; and 4) pending calls. Urgent and less urgent calls are dealt with immediately and are subject to the absolute and relative covering constraints. Pending calls mean that no allocation decision is immediately taken since the patient state is not clear. Such calls are periodically revised.

Even if urgent and less urgent calls are subject to the same covering constraints, their priorities differ. For urgent calls, one or several ambulances closest to the call location are dispatched. All available ambulances are considered, not only those positioned at a fixed location, but also those on their way to a new location. In the latter case, their current position is estimated by the management information system. In the future, the exact position should be transmitted by the GPS. In the case of less urgent call, a different criterion is used. For each possible allocation of an ambulance within r_2 minutes of the call location, the redeployment problem is solved and the allocation yielding the best value of the first term of the objective function (demand covered twice) is selected.

Usually an ambulance is definitively assigned to a call, but in practice the severity of a patient state can lead to modify this rule. More precisely, a vehicle already assigned to a less urgent call may be reassigned to an urgent call if the following two conditions are fulfilled: 1) the reassigned ambulance is the closest ambulance; 2) there exists an ambulance capable of covering the less urgent call within the remaining time. The second condition does not apply when there is no ambulance able to cover the urgent call in time.

3.2 The redeployment problem

We have developed a parallel tabu search heuristic capable of providing high quality solutions. This algorithm is based on a sequential tabu search algorithm previously proposed for a static

ambulance location model Gendreau Laporte and Semet [5]. We now briefly sketch the sequential method and then explain how we parallelized it.

A sequential tabu search heuristic

Gendreau, Laporte and Semet [5] consider a static ambulance location model where the objective is to maximize the total demand covered twice while satisfying the double coverage requirements. The problem is solved using tabu search Glover [6], a local search method which moves at each iteration from a solution to one of its neighbours even if it causes a deterioration of the value of the objective function. To avoid cycling, some solutions features are declared *tabu* for a number of iterations. Several refinements of this general framework have been proposed Glover and Laguna [7], and the method has been applied to a wide variety of contexts Osman and Laporte [12].

The tabu search algorithm developed for the static ambulance location is initialized from a rounded solution of the linear relaxation of the model. At any iteration, the solution may be feasible or infeasible with respect to either covering constraint. Given a solution, a new solution is obtained by moving a set of vehicles between potential location sites. This set is generated by a so-called *aspiration chain*. This chain is constructed through the maximization of a meta-objective function, which takes into account the covering constraints violations, and the model objective function in a hierarchical fashion. More precisely, an ambulance is displaced from a vertex to an *unsaturated* vertex (i.e., a vertex where the maximum number of waiting vehicles has not been attained), and ambulances are moved in order to try to recover first the absolute covering requirements and then the relative covering feasibility. Finally, an attempt is made to increase the value of the objective function by performing non-tabu moves of ambulances. When the best known solution has not been improved for a given number of iterations, a diversification phase is applied. This phase is based on specific rules to initiate the aspiration chain generation process. This tabu approach was tested on randomly generated instances and on instances obtained from the Island of Montreal data. Computational results confirmed that near-optimal solutions could be obtained within modest computing times (up to three minutes on a Sun Sparcstation 1000).

A parallel tabu search heuristic

While the sequential tabu search algorithm is perfectly adequate for the static location problem, a more powerful tool is required for real-time problem solving where life or death decisions must be made in the space of a few seconds several times a day. This is where the power of parallel computing comes into play. To further speed up the decision process we have developed a solution methodology that takes advantage of the available time between consecutive calls by anticipating future decisions on the redeployment of the fleet. More precisely, for each available ambulance, our approach consists of precomputing the relocation decisions associated with the possible assignment of this ambulance to the next incoming call. Thus, when this call actually occurs, an ambulance is assigned according to the rules described previously and the precomputed redeployment scenario is applied. However, when two calls are received in a quick succession, it may still happen with our approach that a fully precomputed solution will not be available. In such a case, no redeployment takes place.

Our parallelization strategy is based on a *master-slave scheme* made possible by the availability of a coarse-grained parallel environment (a network of SUN Ultra Sparc workstations). The master manages mainly two data structures containing for each ambulance available 1) the value of the total demand covered twice and 2) the future positions of the remaining ambulances. Given the current position of the vehicles, these structures can always be initialized. Then, the master distributes the optimization tasks among the slaves in a cyclic way according to increasing order of the total demand covered twice. More precisely, a slave processor is allocated to improve the redeployment strategy associated with an ambulance. The tabu search method is then used to solve the redeployment problem starting from the current best known solution. To control the CPU time allotted to computation of each strategy, the algorithm is run for a given number of iterations. This number is set according to the frequency of calls, which is evaluated as a smoothing average on the inverse of the delay between two consecutive calls. When all strategies have been considered once, a new attempt to improve them is made with a larger number of iterations. Thus, the optimization process is carried out on a full-time basis and the system can provide its best known solution upon request.

To improve the efficiency of the parallel optimization method, the assignment of tasks to slaves takes into account some of the characteristics of the problem. First, even if redeployment decisions are associated with an ambulance, these only depend on the current position of the ambulance. Therefore, when two ambulances are waiting at the same location site, only one relocation strategy has to be computed. Second, since some events are known in advance, (e.g., pending calls, ends of shift, etc), the associated strategies can be computed some minutes before the event occurs. Finally, to increase the CPU time performance of the system, some data updates can be performed in a parallel way. This occurs when a new ambulance enters the system at the beginning of a shift, or when it becomes available after having completed its task. Then, both data structures managed by the master must be updated, which can be achieved by distributing the computation of the relocation strategies among the parallel processors. These data structures store the current solution and the solutions to the subproblems with one fewer ambulance. Each solution is stored as an object containing the ambulance sites and a list of the zones covered by each site. The updating process consists of first stopping the tabu processes and recovering from them improved solutions for the subproblems on which they were working. The relevant solutions are modified accordingly in the data structure. Then, all solutions are adjusted to account for the event that has just occurred. When a new ambulance becomes available, the current solution becomes the solution to the subproblem corresponding to this new vehicle. In all other solutions, the vehicle is simply added at its location. Conversely, when an ambulance is allocated, the solution of its subproblem becomes the new current solution, and the initial solution for each of the subproblems is simply derived by deleting the corresponding vehicle from this new current solution. Note that in both cases the list of zones covered by the sites in each solution must be adjusted to account for the addition or deletion of an ambulance. Finally, the tabu processes for the subproblems are restarted from the new initial subproblem solutions.

We have implemented two additional features to account for the characteristics of ambulance dispatching operations. The first is related to multi-assignment of vehicles to a given call; the second relates to ambulance reassignments. In the event of a multi-assignment, the closest ambulances are dispatched. The redeployment scenarios corresponding to each dispatch are then considered separately and all dispatched ambulances are removed from them. The resulting scenario having the largest demand covered twice is selected and an attempt is made to improve

it by requesting a slave to perform a few tabu search iterations. This is the only situation where a redeployment scenario is computed (or modified) after receiving a call. In the case of an ambulance reassignment, the second best ambulance capable of covering the less urgent call is identified. The redeployment plan associated with this ambulance is computed in priority within the parallelization scheme. However, since less urgent calls have to be covered within r_2 minutes, the ambulance able to cover the less urgent call in the remaining time may change. Thus, following each dispatch, the second best ambulance is always determined. This process is repeated until the ambulance assigned to the less urgent call arrives on the scene.

4. Computational results

The algorithm just described was coded in C++, and PVM was used for the parallel implementation. We used a network of eight Sun Ultra-1/140 workstations (spec int95:5.87; spec fp95:8.38; ram:64M; solaris 7 exploitation system). Six sets of test data were simulated using a real-life call distribution provided by Urgences Santé. More specifically, we worked with the morning call distribution shown in Table 2. The number *c* of calls generated on the six data sets was 120, 120, 130, 130, 140, and 140, yielding an average of 130.

Table 2. Morning call distribution at Urgences Santé

Period *t*	Proportion p_t of morning calls
5:00-6:00	9 %
6:00-7:00	11 %
7:00-8:00	18 %
8:00-9:00	19 %
9:00-10:00	18 %
10:00-11:00	13 %
11:00-12:00	12 %

For each of the seven one hour periods *t* considered, calls were randomly generated according to a Poisson distribution of mean $\lambda_t = c\ p_t$. The spatial distribution of calls was obtained by using 2521 demand points on the Island of Montreal (Figure 1) corresponding to centroids of census tracks, and represented by means of the Universal Transverse Mercator projection. Each of these

points was then weighted by the track population in order to generate more calls in more densely populated areas. The distribution of the four types of call described in section 3.2 was 80%, 3%, 10% and 7%. We assumed all type 2 calls required two ambulances. The number of available ambulances was allowed to vary through the morning. Their speed also varied according to time and each of the three city sectors shown in Figure 1. Realistic speed figures were obtained after discussions with Urgences Santé managers. These values are provided in Table 3. The covering radii were r_1 = 7 minutes and r_2 = 15 minutes and the proportion of the total demand to be covered within r_1 minutes was α = 0.9.

Table 3. Number and speed of ambulances according to time and city sector

Period *t*	Number of ambulances	Speed (km/h)		
		Center	East	West
5h-6h	40	45	50	50
6h-7h	50	40	45	50
7h-8h	60	35	40	50
8h-9h	60	35	40	50
9h-10h	60	35	40	50
10h-11h	58	35	40	50
11h-12h	51	35	40	50

Our tabu search algorithm was run on each of the six data sets. In all simulations covering a morning period of seven hours, all calls were covered within 15 minutes and 98% of urgent calls (types 1 and 2) were covered within 7 minutes, with an average of 3.5 minutes, which is well within the Urgences Santé desired response time of 7 minutes for 90% of urgent calls. Less urgent calls were covered within 9 minutes on average. The algorithm was also very effective since it succeeded in precomputing a strategy in 95% of all cases. The only cases where this was not possible happened when two consecutive calls occurred within 32 seconds of each other. Out of all calls, 38% required at least one ambulance relocation and 99.5% of all relocations involved at most 5 ambulances, with an average of 2.08. This is an interesting result in the sense that our algorithm contained no feature controlling the maximum number of relocated ambulances even if this is one of the constraints taken into account in practice. Finally, to better assess the effectiveness of our heuristic, we extracted 33 scenarios from the simulation data and the tabu

search results were compared with an optimal solution computed by means of CPLEX. On the average, the objective solution value produced by the heuristic was within 2% of the optimal value.

Figure 1. Population distribution on the Island of Montreal

In most parallel computing applications, the effectiveness of the parallelization scheme is measured by comparing the time required by the parallel implementation of the algorithm at hand with that of a sequential implementation, thus yielding a *speedup* index. However, this measure does not really make sense in our context since the algorithm runs continuously in an attempt to improve solutions. Therefore the parallel implementation does not run quicker for a fixed task, but rather performs more work in a set amount of time. One way to derive a meaningful speedup measure is to compare the overall amount of work performed by sequential and parallel variants of our method. This can be measured by the number of tabu search iterations. To this end, we have conducted experiments with one, two, four or eight workstations. The results of these experiments are reported in Table 4. Along with the overall number of iterations performed in each case, we report the average number of applications of the tabu search heuristic to subproblems (scenarios) between two calls. These values indicate how well the real-time system can handle the task of examining the possible scenarios between requests. To correctly interpret these figures, recall that the tabu search heuristic can be applied more than once to a subproblem between two requests. It is interesting to observe that unless eight processors are used, only a fairly small fraction of the subproblems can be examined since in these experiments the average number of unassigned ambulances is around 39. The number of applications of the tabu search heuristic grows sublinearly with the number of processors. This could be expected since some of the applications are interrupted when a new request arrives. The iteration counts indicate that the overall amount of work performed is almost linear in the number of processors, thus confirming the effectiveness of our parallelization scheme.

Table 4. Computational results with 1, 2, 4 and 8 processors

Number of processors	Number of tabu search iterations	Average number of calls to tabu search heuristic between two calls
1	49,005	5.84
2	101,227	11.45
4	197,243	18.47
8	348,392	33.10

5. Conclusion

We have developed a dynamic ambulance dispatching and redeployment system to assist real-time decision-making. The main feature of the system lies in the precomputation of redeployment scenarios that allows immediate decision making when calls are received. High quality solutions as measured by coverage indicators are determined by means of a powerful tabu search engine first developed for a static case. Computational efficiency is achieved through the use of a parallelization strategy which is essential to the success of this application. Computational results based on real data show that the proposed system can effectively solve real-life instances such as those arising in Montreal.

Acknowledgements

This work was partly supported by the Fond de la Recherche en Santé du Québec (FRSQ) (grant 980861) and by the Canadian Natural Sciences and Engineering Council (grants OGP0038816, OGP0039682 and OGP0184123). This support is gratefully acknowledged. Thanks are also due to Alain David, François Guertin and Hervé Lomeret for their help with programming, and to two reviewers for their valuable comments.

References

[1] M.O. Ball, F.L.Lin. A Reliability model applied to emergency service vehicle location. Operations Research 41 (1993) 18-36.

[2] L. Brotcorne, L. Farand, G. Laporte, F. Semet. Impacts des nouvelles technologies sur la gestion des systèmes de véhicules d'urgence. Ruptures 6 (1999) 149-168.

[3] C. Desrosiers, G. Thibault. Réingénierie des processus: cas vécu à Urgences Santé. Presented at Séminaire de la Fédération de l'Informatique du Québec. 1996.

[4] M. Gendreau, F. Guertin, J.-Y. Potvin, E. Taillard Parallel tabu search for real-time vehicle routing and dispatching. Transportation Science 33 (1999) 381-390.

[5] M. Gendreau, G. Laporte, F. Semet Solving an ambulance location model by tabu search. Location Science 5 (1997) 75-88.

[6] F.Glover Future paths for integer programming and links to artificial intelligence. Computers & Operations Research 13 .(1986) 533-549.

[7] F. Glover, M. Laguna. *Tabu Search*. Kluwer, Boston, 1997.

[8] K. Hogan, C.S. ReVelle. Concept and applications of backup coverage. Management Science 32 (1986) 1434-1444.

[9] M.B. Mandell. Covering models for two tiered emergency medical systems. Location Science 6 (1998) 355-368.

[10] V. Marianov, C.S. ReVelle. The queueing maximal availability location problem: A model for the sitting of emergency vehicles. European Journal of Operational Research 93 (1996) 110-120.

[11] V. Marianov, C.S. ReVelle. Siting emergency services. In: *Facility Location : A Survey of Applications and Methods*, Z. Drezner (ed.) Springer-Verlag New York, (15) 199-223.

[12] I.H. Osman, G. Laporte. Metaheuristics: A Bibliography. Annals of Operations Research 63 (1996) 513-628.

[13] C.S. ReVelle, K. Hogan. The maximum availability location problem. Transportation Science 23, (1989) 192-200.