Gest. Prod., São Carlos, v. 23, n. 4, p. 828-841, 2016
http://dx.doi.org/10.1590/0104-530X2100-16

Positioning of ambulances of the SAMU system by Integer Programming and Queueing Theory

Posicionamento de ambulâncias do SAMU através de Programação Inteira e Teoria de Filas

Bruno Barreto[1]
Fernando Alexandrino[1]
Ormeu Coelho[1]

Abstract: The configuration of emergency services logistic networks for is a paramount strategic issue since small deviations may lead to death of users. From this premise, the work proposes new alternatives for positioning the SAMU ambulance system in the city of Duque de Caxias / RJ, which are able to reduce the service response time. These ambulances repositioning proposals were built in two stages: at first two models of Integer Programming were used in order to obtaining solutions that maximizes coverage. Then, the Hypercube Model was applied for evaluating the server's availability under randomness, as well as other relevant performance indicators, such as average time response, and servers' workloads.

Keywords: Facility location; Hypercube Model; Emergency services.

Resumo: *A configuração de redes logísticas para serviços de emergência é questão estratégica de imensa importância, visto que pequenas variações no tempo de resposta podem implicar na morte do solicitante. Partindo dessa premissa, o trabalho propõe novas alternativas de posicionamento para as ambulâncias do sistema SAMU na cidade de Duque de Caxias, RJ, capazes de reduzir o tempo de resposta do serviço. Essas propostas de reposicionamento das ambulâncias foram construídas em duas etapas: na primeira, dois modelos de Programação Inteira foram aplicados para se obter soluções que provejam maior cobertura à população. Posteriormente, o Modelo do Hipercubo foi empregado para avaliar a disponibilidade dos servidores, dentre outros indicadores de desempenho relevantes, como o tempo médio de resposta e a taxa de ocupação das ambulâncias.*

Palavras-chave: *Localização de facilidades; Modelo do Hipercubo; Serviços de emergência.*

1 Introduction

When designing a logistic network it is essential to pay attention to the particularities inherent to an operation in the private or public sector. While the former searches for a network in which the products flow between points of supply and demand based on the inherent cost/profits, in the public sector the central issue is to optimize any function that measures the availability of the service for a given population (Ghiani et al., 2004).

Within the public administration, the three main objectives to be considered during the design of a network are: budget; operational cost reduction; and the increasing of the service level (Ghiani et al., 2004). Particularly in healthcare, one of the logistical problems of greatest interest is the ambulance location that will attend emergency calls. The location of these facilities is very sensitive to the required service level, mainly characterized by the service response time. A poor coverage may imply the death of the user.

The response time or, as will be discussed later, the coverage limit set by US law is 10 minutes maximum for urban areas within a service level of 95%, and may be extended to 30 minutes for rural areas (Ball & Lin, 1993). In London, 95% of the requests must be met within 14 minutes (Galvão et al., 2003a). In Montreal, the maximum response time must be less than or equal to 10 minutes for 70% of calls (Gendreau et al., 2001). However, according to Takeda et al. (2004), in Brazil there is no legislation to determine an upper limit for the response time of this kind of service.

In this paper, the repositioning of ambulances of the Mobile Emergency Service (SAMU), in Duque

[1] Departamento de Engenharia de Produção, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca – CEFET/RJ, CEP 20271-110, Rio de Janeiro, RJ, Brazil, e-mail: brunobarreto7@gmail.com; fernando.hnd@gmail.com; ormeucoelho@gmail.com

Received Mar. 1, 2015 - Accepted June 12, 2015

Financial support: None.

de Caxias/RJ, is analyzed, using the combination of Integer Programming techniques (IP) and Queuing Theory. Although there is extensive use of deterministic optimization models in such problems as, for example, Schmid & Doerner (2010), Iannoni et al. (2009) and Gendreau et al. (1997), they do not properly evaluate the congestion effects on the servers. The current work is characterized by the use of a queuing model in order to evaluate the solutions generated by models of IP, used to locate SAMU ambulances. The use of Hypercube Model (HM) in congested systems can represent aspects fleeing to the above deterministic formulations (Larson, 1974).

This study also aims to evaluate the possibility of increase the number of users served within an acceptable time limit by just reposition the servers in locations that already have a minimum infrastructure. This idea is aligned with the concept of "decentralized basis" presented in the Ministerial Order No. 2657 of the Ministry of Health, 16 December 2004, which allows for decentralized bases that act as outposts for ambulances and their teams, thus ensuring a quality response time to users of SAMU (Brasil, 2013).

This paper was divided into five sections. Section 2 presents a brief literature review on the facility location models for emergency services coverage. The current performance of system was analyzed through coverage indicators, as shown in Section 3. Section 4 discusses the application of two IP models surveyed to the data of the SAMU – Duque de Caxias/RJ, and the obtained results are presented in Section 5. It also reports the proposals for positioning the SAMU ambulances in order to maximize service coverage. In this sense, the HM is used to assess part of the best solutions obtained by one optimization package for the IP. Conclusions and perspectives for future research are presented in Section 6.

2 Literature review

The Set Covering Problem (SCP) was one of the first discrete models for facility location used in emergency services (Toregas et al., 1971). It considers a coverage constraint expressed by the maximum travel time or distance between facility/server and client in a logistic network. Such separation measure is sometimes referred as "critical distance", *S*. The problem is defined over a network in which *I* is the set of demand nodes, and *J* is the set of candidates points to the allocation of servers. The node i ($i \in I$) is considered covered by the service if and only if the separation (measured in units of distance/time) between client node i and the nearest ambulance, located at some node ($j \in J$), is less than or equal *S*.

Unlikely the SCP, the Maximum Coverage Problem proposed by Church & Revelle (1974) seeks to maximize the population covered within the "critical distance", *S*, given a predefined number of facilities, *p*. In this model, the number of number of facilities/servers is determined exogenously by the existence of limited budget or managerial restrictions.

The TEAM model (Tandem Equipment Allocation Model) assumes the existence of two separate servers, each one with its respective critical distance. This premise is very common in emergency services, as there are servers with different equipment, which are able to attend different events (Schilling et al., 1979).

There are situations in which the service provided by the basic units may also be provided by the advanced ones. So, a higher coverage can be achieved by allowing the servers to be positioned independently. Such adaptation was presented by Schilling et al. (1979) in the model known as FLEET (Facility-location, Equipment-emplacement Technique). This model requires that each client node is simultaneously covered by primary servers (or base) and special (or advanced). The sets $N_i^p = \{j \in J | d_{ji} \leq S^p\}$ and $N_i^s = \{j \in J | d_{ji} \leq S^s\}$ contain the nodes where the allocation of a server, primary and special, in this order, allows the coverage of node i. d_{ji} is the "distance" between nodes j and i, through some shortest path in the network, and S^p and S^s are the "critical distances" to ensure coverage by basic and special servers, respectively. The formulation also considers the parameters and variables below:

- a_i = population of node i;
- P^p = number of primary servers;
- P^s = number of special servers;
- P^z = number of facilities to be installed.

The decision variables are:

- $x_j^p = \begin{cases} 1, & \text{if a primary ambulance is allocated to node } j; \\ 0, & \text{otherwise.} \end{cases}$
- $x_j^s = \begin{cases} 1, & \text{if a special ambulance is allocated to node } j; \\ 0, & \text{otherwise.} \end{cases}$
- $z_j = \begin{cases} 1, & \text{if a facility is opened at node } j; \\ 0, & \text{otherwise.} \end{cases}$
- $y_i = \begin{cases} 1, & \text{if node } i \text{ is covered;} \\ 0, & \text{otherwise.} \end{cases}$

The mathematical formulation of FLEET is then given by Expressions 1-11:

$$\max Z = \sum_{i \in I} a_i y_i \tag{1}$$

s.t.:

$$\sum_{j \in N_i^p} x_j^p \geq y_i, \qquad \forall i \in I \tag{2}$$

$$\sum_{j \in N_i^s} x_j^s \geq y_i, \qquad \forall i \in I \tag{3}$$

$$\sum_{j \in J} x_j^p = P^p, \tag{4}$$

$$\sum_{j \in J} x_j^s = P^s, \tag{5}$$

$$\sum_{j \in J_N} z_j = P^z, \tag{6}$$

$$x_j^p \leq z_j, \qquad \forall j \in J_N \tag{7}$$

$$x_j^s \leq z_j, \qquad \forall j \in J_N \tag{8}$$

$$y_i \in \{0,1\}, \qquad \forall i \in I \tag{9}$$

$$z_j \in \{0,1\}, \qquad \forall j \in J_N \tag{10}$$

$$x_j^p, x_j^s \in \{0,1\}, \qquad \forall j \in J \tag{11}$$

The objective Function 1 seeks to maximize the covered population, while the Constraints 2 and 3 compute the coverage of node *i* only when it is covered by at least one basic and one advanced ambulance, respectively. The Equations 4 and 5 defines the availability of servers of each kind. The set of nodes that are able to receive a facility is J_N, $J_N \subset J$, and exactly P^z facilities must be installed, as indicated in (6). The Constraints 7 and 8 ensure that ambulances are allocated only to nodes which have a facility.

One of the first stochastic approaches to locate facilities/emergency call servers was the Maximum Availability Location Problem (MALP) proposed by Revelle & Hogan (1989). MALP seeks to model the uncertainty inherent the demand by simplifying assumptions. *P* servers must be located in order to maximize the population covered within *S*, with reliability θ (Galvão et al., 2003b). It assumes that servers operate at the same busy fraction, ρ, which is formally defined in (12).

$$\rho = \frac{\bar{t} \sum_{i \in I} \lambda_i}{24 \sum_{j \in J} y_j} = \frac{\bar{t} \sum_{i \in I} \lambda_i}{24P} \tag{12}$$

such that:

- λ_i = arrival rate of calls at node $i \in I$;
- $\bar{t}$ = average duration of a call (in hours);
- P = number of servers.

The single group of decision variables is y_j, such that:

- $y_j = \begin{cases} 1, \text{if a server is located at node } j; \\ 0, \text{otherwise.} \end{cases}$

The minimum number of servers, *b*, needed to cover a given node using the confidence level θ can be obtained from (12). This is done by computing the probability of having at least one ambulance available to answer a call within "critical distance", *S*, given the arrival rate of calls (Revelle & Hogan, 1989). Expression 13 calculates this probability, such that c_{ji} are the coefficients of the binary matrix, whose value is 1 if $d_{ji} \leq S$, and 0, otherwise.

$$\begin{aligned} &P(\text{if at least one server is within critical distance } S) \geq \theta \\ &= \left[1 - P(\text{none server is within critical distance } S)\right] \geq \theta \\ &= 1 - \rho^{\sum_{j \in J} c_{ji} y_j} \geq \theta \end{aligned} \tag{13}$$

The sum $\sum_{j \in J} c_{ji} y_j$ defines the number of servers available within "critical distance" *S* from a given demand node $i \in I$. In order to cover a node with reliability θ, there must be at least *b* servers able to answer this a call from this node. Computing the logarithms of both members in (13) it becomes $\sum_{j \in J} c_{ji} y_j \geq b$, where $b = \frac{\log(1-\theta)}{\log \rho}$. That is, given the busy fraction ρ, we are able to compute the number of required facilities for guaranteeing coverage with reliability θ.

The MALP variables are:

- $y_{ik} = \begin{cases} 1, \text{if node i is covered by at least } k \text{ ambulances;} \\ 0, \text{otherwise.} \end{cases}$
- $x_j = \begin{cases} 1, \text{if an ambulance is located at node } j; \\ 0, \text{otherwise.} \end{cases}$

The mathematical formulation of MALP is given by the Expressions 14-18:

$$\max Z = \sum_{i \in I} \lambda_i y_{ib} \tag{14}$$

s.t.:

$$\sum_{k=1}^{b} y_{ik} \leq \sum_{j \in J} c_{ji} x_j, \qquad \forall i \in I \tag{15}$$

$$y_{ik} \leq y_{i(k-1)}, \qquad \forall i \in I,\ k = 2, \ldots, b \tag{16}$$

$$\sum_{j \in J} x_j = P \tag{17}$$

$$x_j, y_{ik} \in \{0,1\} \qquad \forall i \in I,\ \forall j \in J,\ k = 2, \ldots, b \tag{18}$$

The objective function (14) maximizes the population covered within critical distance *S* and reliability θ, that is, only the nodes covered by $k = b$ ambulances are computed. The left-hand side in (15) counts the number of servers within the "critical distance" *S* from the demand node *i*, thus ensuring coverage when there are *b* servers. Moreover, if k

ambulances cover node *i*, then it is true that *i* is also covered by $k-1$ servers, as expressed by Constraint 16. Finally, Constraint 17 defines the number of ambulances.

Although MALP considers the probability that the server is available when a call arrives, there is no guarantee that the uncertainty related to the arrival process has been well modeled, which can significantly affect the reliability of answering calls, causing the formation of queues and increasing the time of service. Furthermore, the assumption that the servers have the same busy fraction is unlikely in a real situation. Batta et al. (1989) also support this statement pointing causes as the disproportionate distribution of demand along the served region, and the dispatching policy that can prioritize certain servers, thus unbalancing the fraction of time they are occupied. Despite the existence of models that consider a specific busy fraction for each server, it is difficult to infer their values as they are output from the positioning computed by the location model. Brotcorne et al. (2003) suggest the usage of simulation or queuing theory to obtain them.

Therefore, given a solution obtained by one of the models just discussed above, it is interesting to evaluate it through indicators which are influenced by uncertainty (such as average answer time, average number of waiting users, among others). In this paper this evaluation is done by the Hypercube Model proposed by Larson (1974) and commonly used to model systems in which servers are dispatched to the customers in order to provide a service, and demand is geographically distributed by discrete atoms.

The area under study is divided into *I* geographic atoms (nodes) and the arrival of calls from atom *i* is a Markovian process with rate λ_i. In order to answer these calls, the system has *N* servers distributed among the atoms, whose response time is exponentially distributed with attendance rate μ_n (Chiyoshi et al., 2000). Every server can be in one of two states, free (0) or busy (1), and the combination of the states of all servers results in the system's state. For example, for a system with three servers, the {001} state indicates server #1 is occupied, while servers #2 and #3 are free. Hence, the number of possible states is 2^N.

We assume only one server is dispatched to the same call, and there is a priority order among servers for answering a call originated from atom *i*. If the server of higher priority is busy, the second is dispatched, and so on until the last. In state {11...1} all servers are busy, and any new call must to wait in a queue, according to FCFS (First Come, First Served) policy. So, besides the 2^N states mentioned previously, there also are the states in which u calls are in the system, such that $u \geq N+1 : \{S_{N+1}\}, \{S_{N+2}\}, \{S_{N+3}\}, \ldots$ These states may be represented by the vertices of a hypercube, which has inspired the model's name (Chiyoshi et al., 2000). Figure 1 illustrates a system with $N = 3$ servers and queue limited to *l* calls.

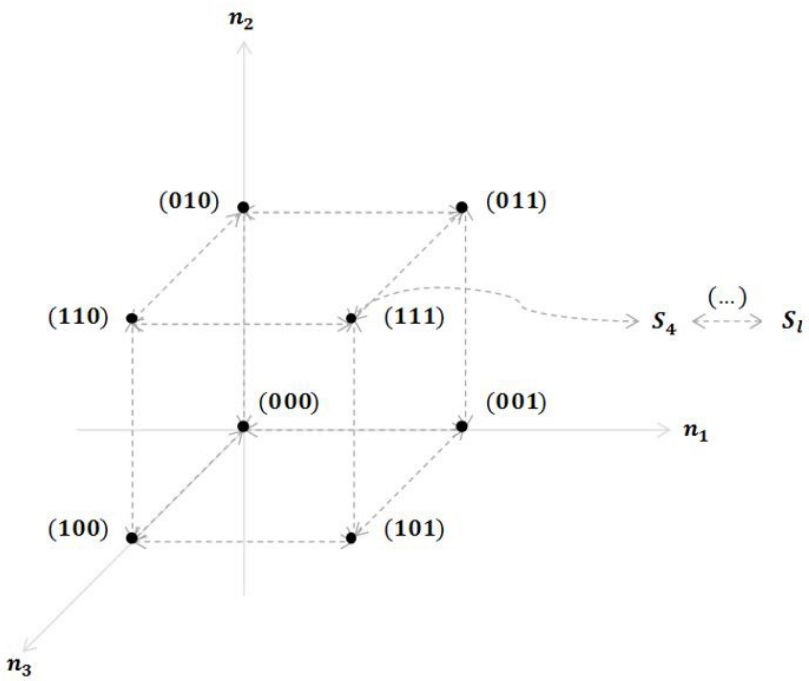


Figure 1. Possible states for three servers and finite queue. Source: Elaborated by the authors.

Larson (1974) defines two transition classes in a Hypercube: upward, in which the server goes from free to busy, and downward when the server changes from busy to free. It is assumed transitions only occur between adjacent vertices of the Hypercube, and the rate at which the system enters a particular state is equal to the rate it leaves this state, so the model's equilibrium equations are built. By taking an example with $N = 3$ servers and $I = 3$ atoms and assuming p_B as the probability of the system to be in state B the equilibrium equation around $B = \{000\}$ is then expressed in Equation 19, where λ is the total call rate of the system:

$$\lambda p_{000} = \mu_1 p_{001} + \mu_2 p_{010} + \mu_3 p_{100} \quad (19)$$

In Equation 19, $\lambda = \lambda_1 + \lambda_2 + \lambda_3$, since from the state {000} the system can enter the states {001}, {010} or {100} by receiving a call originating from atom #1, #2 or #3, respectively. The right-hand side indicates the possibility of reaching the {000} state from the answering of any call when the system is in one of the states {001}, {010} or {100}, which occur, respectively, with rates μ_i, $i = 1,2,3$. Considering that the server n is located at atom $i = n$ (so, it is preferred to serve itself), the state equation {001} is constructed in a similar way:

$$(\lambda + \mu_1) p_{001} = \lambda_1 p_{000} + \mu_2 p_{011} + \mu_3 p_{101} \quad (20)$$

State {111} is obtained by receiving a call when the system is in the states {011}, {101} or {110}, regardless the preference order, and also by the answering of any call when the system is in state S_4 in which there are three users being in service and one in the queue. Once a server completes the service, it becomes available and it is dispatched to answer the call that had been queued, bringing the system back to the state {111}. Thus, the equation for the state {111} is in (20):

$$(\lambda + \mu) p_{111} = \lambda p_{011} + \lambda p_{101} + \lambda p_{110} + \mu p_4 \quad (21)$$

Given the system's equilibrium condition, the transition rates between the states {111} and S_4 shall be equal, i.e., $\lambda p_{111} = \mu p_4$. Likewise, transitions between states and $\{S_l\}$ e $\{S_{l+1}\}$, to $l \geq N$ are equal and equivalent to $\rho = \frac{\lambda}{\mu}$. The transition equations when all servers are busy forms a geometric progression, so $p_{111} + p_4 + p_5 + \ldots = p_{111} / (1 - \rho)$. As the sum of the probabilities of all states of the system is equal to 1, it is possible to obtain the normalization Equation 22, resulting in a linear system that has a unique solution with 2^N equations.

$$p_{000} + p_{001} + p_{010} + \ldots + p_{111} / (1 - \rho) = 1 \quad (22)$$

By solving this system, different performance metrics can be computed, as the average response time, the busy fraction and the probability of queuing, which allows one to analyze how the geographic arrangement react when subjected to stochastic demand.

Among the studies that used the Hypercube Model it is worth also mentioning the works of Larson (1975), Brandeau & Larson (1986), Galvão et al. (2003b), Takeda et al. (2004) and Souza et al. (2013).

3 Field research description

Duque de Caxias is a city of the metropolitan region of Rio de Janeiro, which according to the IBGE census had 855,048 inhabitants by 2010 (IBGE, 2014). To serve this population, SAMU has seven basic life support (BLS) and two advanced life support ambulances (ALS). As in many Brazilian cities, the Health Department of Duque de Caxias does not have any computational tool for locating these ambulances, which is made empirically most times. The operational coordination of SAMU is located at Hospital Dr. Moacir Rodrigues do Carmo, where the two advanced units are positioned. Table 1 shows the current distribution of ambulances in the city.

3.1 Logistical network construction

At first, all public hospitals, health care centers and UPAs (emergency care units) in the city were considered as candidate locations to install a facility, resulting in the set J defined in the Formulations 1-11 and 14-18. In addition, the geographical area of the

Table 1. Ambulance distribution existing by the time of this research.

Facility	Address	ALS	BLS
Hospital Municipal Dr. Moacir R. do Carmo	3200 Washington Luiz Hwy., Beira Mar	2	2
Posto Médico Sanitário de Campos Elíseos	333 Actura St., Campos Elíseos	0	1
Posto Médico Sanitário Parque Equitativa	Automóvel Clube Ave., Parque Equitativa	0	1
Posto Médico Sanitário do Pilar	Carlos Alves St., Pilar	0	1
Posto Médico Sanitário Saracuruna	Presidente Roosevelt Ave., Saracuruna	0	1
Posto Médico Sanitário de Xerém	Nóbrega Ribeiro St., Xerém	0	1

Source: Elaborated by the authors.

Duque de Caxias was divided into sub regions as small as possible given the available data. In each of them, there was chosen arbitrarily a "center", where all demand is assumed to be concentrated, originating the set I, in such a way that $J \subset I$. Table 2 describes the vertices I considered in the study, such that the shaded nodes are those belonging to set J, while the others belong only to set I. The location of each atoms in terms of latitude (LAT) and longitude (LONG) can be found in Table 3, as well as their population a_i, which was obtained based on IBGE census in 2010 (IBGE, 2014).

At last, a directed network $G = (V, A)$ was built, so that the arcs in the set A represent the possibilities of treks between pairs of vertices (through streets, avenues or alleys). These arcs are valued with the travel times among the pairs of nodes in V, $V = I$. It is assumed that when attending a call the ambulance uses the shortest route between its current location and user's location. The travel times among facilities and the demand nodes were estimated based on the values corresponding to peak times, between 5pm and 7pm, by consulting the API (Application Programming Interface) by Google Maps. Table 4 shows the travel times (in minutes), where the shaded values correspond to those which $c_{ji} = 1$, while for others, $c_{ji} = 0$.

3.2 Current arragement analysis

According to Bertelli et al. (1999), the highest survival frequency of cardiac arrest victims occurs when the resuscitation maneuvers are carried out within 8 minutes. This parameter was used as response time limit for the advanced type servers, i.e., $S^S = 8$ minutes. For basic ambulances, it was adopted $S^p = 12$ minutes. Considering the positioning of ambulances and the estimated travel times, we used the HM to evaluate the performance parameters of the current logistic arrangement.

According to SAMU's operational database, 17,862 calls were received between January and June 2013. Using the assumptions adopted in HM, the number of arrivals during the time interval t follows a Poisson distribution with mean λ if and only if the time between arrivals is exponentially distributed with mean $1/\lambda$ (Taha, 2008). In order to verify this assumption, the mean interval times between successive calls were taken over 21 days, from the above mentioned database, as shown in Table 5.

Table 2. Network nodes.

Address	Node	Address	Node
Pam 404 Doutor Fernando Gil	01	Duque de Caxias, CEP 25250-400	25
Posto de Saúde Alaide Cunha	02	Duque de Caxias, CEP 25271-350	26
Duque de Caxias, CEP 25235-460	03	Duque de Caxias, CEP 25036-600	27
Duque de Caxias, CEP 25015-415	04	Duque de Caxias, CEP 25272-410	28
Duque de Caxias, CEP 25267-390	05	Duque de Caxias, CEP 25265-232	29
Posto Médico Sanitário de Campos Elíseos	06	Hospital Municipal Dr. Moacir R. do Carmo	30
Duque de Caxias, CEP 25220-570	07	Duque de Caxias, CEP 25240-650	31
Duque de Caxias, CEP 25245-230	08	Duque de Caxias, CEP 25046-380	32
Centro Municipal de Saúde de Duque de Caxias	09	Posto de Saúde Sarapuí	33
Hospital Infantil Ismélia Silveira	10	UPA Sarapuí	34
UPA Infantil Walter Garcia	11	Duque de Caxias, CEP 25025-300	35
UPA Duque de Caxias	12	Posto Médico Sanitário do Pilar	36
Duque de Caxias, CEP 25251-100	13	Posto Médico Sanitário Santa Cruz da Serra	37
Duque de Caxias, CEP 25243-150	14	Duque de Caxias, CEP 25271-430	38
Duque de Caxias, CEP 25237-030	15	Duque de Caxias, CEP 25040-060	39
Posto Médico Sanitário Parque Equitativa	16	Duque de Caxias, CEP 25045-040	40
Duque de Caxias, CEP 25060-190	17	Posto Médico Sanitário Saracuruna	41
Duque de Caxias, CEP 25231-180	18	Duque de Caxias, CEP 25270-450	42
Posto Médico Sanitário Dr. Jorge R. Pereira	19	Duque de Caxias, CEP 25030-180	43
Posto de Saúde Doutor José de Freitas	20	Duque de Caxias, CEP 25040-610	44
Posto de Saúde Edna Salles	21	Duque de Caxias, CEP 25065-162	45
Posto de Saúde José Camilo dos Santos	22	Hospital Municipal Maternidade de Xerém	46
Hospital Estadual Adão Pereira Nunes	23	Unidade Pré-Hospitalar Álvaro Figueira	47
Duque de Caxias, CEP 25250-130	24	Posto Médico Sanitário de Xerém	48

Source: Elaborated by the authors.

Table 3. Population distribution.

Node	District	LAT	LONG	a_i	Node	District	LAT	LONG	a_i
01	25 de Agosto	-22.793	-43.299	7,071	25	Mantiquira	-22.596	-43.302	10,616
02	25 de Agosto	-22.786	-43.297	7,071	26	Meio da Serra	-22.626	-43.206	2,344
03	Amapá	-22.676	-43.357	6,477	27	Olavo Bilac	-22.766	-43.328	34,770
04	Bar dos Cavaleiros	-22.795	-43.326	41,209	28	Parada Angélica	-22.629	-43.210	14,458
05	Barro Branco	-22.638	-43.244	15,700	29	Parada Morabi	-22.657	-43.230	4,444
06	Campos Elíseos	-22.660	-43.250	19,622	30	Parque Duque	-22.799	-43.289	44,983
07	Cangulo	-22.688	-43.236	13,053	31	Parque Eldorado	-22.635	-43.307	8,161
08	Capivari	-22.647	-43.328	1,489	32	Parque Fluminense	-22.725	-43.319	34,969
09	Centro	-22.787	-43.308	6,756	33	Parque Sarapuí	-22.751	-43.296	1,009
10	Centro	-22.788	-43.311	6,756	34	Parque Sarapuí	-22.751	-43.299	1,009
11	Centro	-22.793	-43.307	6,756	35	Periquitos	-22.779	-43.324	17,898
12	Centro	-22.786	-43.325	6,756	36	Pilar	-22.711	-43.306	33,525
13	Chácaras Arcampo	-22.657	-43.274	14,120	37	Santa Cruz da Serra	-22.645	-43.274	25,698
14	Chácaras Rio-Petrópolis	-22.665	-43.315	14,085	38	Santa Lúcia	-22.625	-43.210	16,732
15	Cidade dos Meninos	-22.630	-43.222	2,460	39	Santo Antônio	-22.745	-43.317	11,420
16	Cidade Parque Paulista	-22.635	-43.263	33,501	40	São Bento	-22.728	-43.305	22,062
17	Doutor Laureano	-22.764	-43.299	43,996	41	Saracuruna	-22.676	-43.254	46,660
18	Figueira	-22.680	-43.298	16,520	42	Taquara	-22.627	-43.236	12,191
19	Imbariê	-22.636	-43.217	34,332	43	Vila Centenário	-22.774	-43.313	21,922
20	Jardim Anhangá	-22.637	-43.231	12,867	44	Vila São José	-22.742	-43.317	31,009
21	Jardim Gramacho	-22.761	-43.278	53,731	45	Vila São Luís	-22.773	-43.298	30,420
22	Jardim Primavera	-22.695	-43.261	20,915	46	Xerém	-22.599	-43.302	7,466
23	Jardim Primavera	-22.670	-43.279	20,915	47	Xerém	-22.600	-43.292	7,466
24	Lamarão	-22.598	-43.293	192	48	Xerém	-22.601	-43.292	7,466

Source: Elaborated by the authors.

Using the Kolmogorov-Smirnov test for average with significance level $\alpha = 0.05$, it was obtained a *p*-value = 0.9, which indicates that one cannot reject the null hypothesis that the interval between successive arrivals follows a negative exponential distribution and. Hence, the number of calls in *t* is a Poisson process with mean $\hat{\lambda} = 1/2.39 = 0.42$ calls/hour. Given the unavailability of data disaggregated by geographic atoms, the call rate of each atom (λ_i) was estimated using the same approach adopted by Takeda et al. (2004). The authors suggest to approximate λ_i by the multiplicative product between p_i (probability of a call is originated from atom *i*, i.e., the calls relative percentage of calls from that atom) and $\hat{\lambda}$ (the total call rate of the system). Table 6 shows the obtained estimates.

The total operational time is defined as the sum of: vehicle preparation time, travel to the demand node, victim attendance and return. The average over the 21 days was analyzed and the standard deviations (in minutes) are shown at Table 7 and distinguished by each server, where the first two are advanced units. According to Takeda et al. (2004), when deviations have the same magnitude order as the average, as this particular case, it can be inferred that the distribution is approximately exponential. This hypothesis was confirmed by the Kolmogorov-Smirnov test with $\alpha = 0.05$, from which was computed a *p*-value greater than 0.7. Therefore, the service rates $\mu^s = 60/77 = 0.78$ calls/hour and $\mu^p = 60/75 = 0.80$ calls/hour were computed for the ASL and BSL units, respectively. At last, the preference order matrix was created, listing for each atom *i* the closest servers (in terms of travel time), in ascending order, with no distinction between the vehicle's type (basic or advanced). The server in the first column of row *i* is then the preferred, and the others are used as backups.

The system allows queuing and even though it is unrestricted in practice, it was adopted a capacity of nine users (the number of servers) in order to calculate the probability of arriving a call when the queue is so large that it would be considered lost. The probability of receiving a tenth call when the queue already is at full capacity can be calculated as $\rho^{10} p_{1...1}$. In the simulation of the current logistic arrangement, the probability of this event was 0.03%, while in the proposed scenarios it was less than 10^6, which indicates that this limitation does not bring significant changes to the performance indicators of HM.

The average travel times in this system $(\bar{T})$ and to each atom $(\bar{T}_i)$ are computed by expressions (23) and (24), respectively (Chiyoshi et al., 2000):

Table 4. Travel time t_{ji} (minutes).

Node	01	02	06	09	10	11	12	16	19	20	21	22	23	30	33	34	36	37	41	46	47	48
01	2	4	24	3	8	5	13	23	24	23	12	20	16	7	11	12	16	22	23	23	22	22
02	4	4	22	5	8	7	10	21	22	21	10	18	14	9	9	10	14	20	21	22	20	20
03	27	26	28	25	23	24	27	23	28	27	27	29	20	28	19	19	15	22	29	16	16	16
04	10	11	27	8	6	8	4	26	27	26	16	23	19	13	15	16	19	25	26	26	25	25
05	27	26	9	29	29	30	33	6	10	5	22	22	13	23	24	25	23	8	14	17	15	15
06	18	18	29	20	20	21	25	13	18	18	14	11	10	15	15	16	15	12	5	13	12	11
07	27	26	30	29	29	30	34	24	30	29	22	10	23	23	24	25	23	23	10	25	23	23
08	26	26	20	28	29	29	33	15	20	19	21	21	13	23	23	25	19	14	21	13	13	12
09	4	5	26	3	8	7	10	25	26	25	14	22	19	9	13	13	19	25	25	26	25	24
10	7	8	28	5	1	4	6	27	28	27	16	24	21	12	13	13	19	26	27	28	26	26
11	9	11	25	7	4	2	9	24	25	24	14	21	18	10	13	14	17	23	24	25	23	23
12	9	9	29	8	5	8	7	28	29	28	17	25	21	14	15	15	22	27	28	28	27	27
13	18	18	12	20	20	21	25	7	12	11	14	14	4	15	15	17	15	3	13	12	10	10
14	25	25	19	27	27	28	32	14	19	18	21	20	11	22	22	23	18	13	20	15	13	13
15	26	26	8	28	28	29	33	9	2	6	22	20	16	23	23	24	23	11	13	20	18	18
16	25	24	11	27	27	28	32	5	10	6	20	20	12	21	22	23	21	6	16	15	14	13
17	12	8	21	9	11	11	13	20	21	20	9	17	14	10	6	6	14	19	20	21	19	19
18	21	21	15	23	25	26	29	14	15	14	17	16	7	18	18	19	13	13	16	15	13	13
19	25	24	6	27	26	27	31	11	4	5	20	18	18	21	22	23	21	13	11	19	18	18
20	25	24	6	27	26	27	31	7	5	4	20	18	14	21	22	23	21	9	11	18	16	16
21	12	11	21	14	14	14	19	20	22	21	3	17	14	8	8	9	14	20	21	21	20	20
22	21	20	24	23	22	23	27	20	24	23	16	9	16	17	18	19	16	19	8	21	19	19
23	17	16	9	19	19	20	24	11	9	8	13	13	16	14	14	15	13	7	11	13	11	11
24	25	25	19	27	27	28	32	13	19	19	21	20	12	22	22	23	22	13	20	3	2	2
25	26	26	21	29	28	29	33	14	21	20	22	22	13	23	24	25	23	14	21	2	3	3
26	27	26	8	29	28	29	33	12	4	8	22	20	19	23	24	25	23	14	13	21	20	20
27	15	14	33	14	11	14	8	32	33	32	22	29	25	20	14	14	21	31	32	32	31	31
28	26	25	8	28	28	29	33	12	3	7	21	19	19	22	23	24	22	14	13	20	19	19
29	23	22	5	25	25	26	30	13	6	7	18	16	15	20	20	21	19	15	10	18	16	16
30	5	8	22	7	8	8	13	21	22	21	11	18	15	6	10	11	14	20	21	22	20	20
31	27	27	22	29	29	30	34	13	22	18	23	23	14	24	25	26	25	12	22	8	7	7
32	15	15	22	14	12	13	16	21	23	22	16	23	15	17	8	8	8	20	23	21	20	20
33	13	13	18	11	10	10	14	17	18	17	8	14	11	9	2	2	11	17	18	18	17	16
34	13	13	20	11	9	10	14	19	21	20	9	16	13	10	2	1	13	19	20	20	19	19
35	11	10	30	10	7	10	4	29	30	29	18	26	23	16	16	16	24	29	30	30	29	28
36	20	20	18	19	17	17	21	16	18	17	20	20	11	21	13	13	5	16	19	17	16	16
37	21	20	13	23	22	24	27	5	13	11	16	16	7	17	18	19	17	2	15	12	11	10
38	27	27	9	29	29	30	34	11	4	9	22	20	19	24	24	25	23	14	14	22	21	21
39	14	13	25	13	10	11	15	24	25	24	14	21	18	15	7	7	14	24	24	25	24	23
40	16	15	25	14	13	13	16	23	25	24	16	23	17	17	8	8	10	22	25	23	22	22
41	21	20	21	23	23	24	28	15	21	20	16	8	13	17	18	19	17	14	6	15	14	14
42	28	28	13	30	30	31	35	7	8	7	24	23	15	25	25	26	25	10	17	18	17	17
43	9	6	27	6	5	6	7	26	27	26	15	23	20	13	11	11	18	25	26	27	25	25
44	15	15	28	13	12	13	16	26	28	27	15	22	20	16	8	8	13	25	26	27	25	25
45	9	5	21	6	8	9	11	20	21	20	9	17	13	9	8	9	13	19	20	21	19	19
46	26	26	20	28	28	29	33	14	20	20	21	21	13	23	23	24	22	13	21	1	3	3
47	25	24	19	27	27	28	32	13	19	18	20	20	11	21	22	23	21	13	19	3	4	1
48	24	24	18	26	26	27	31	12	18	18	20	19	11	21	21	22	20	11	19	3	1	4

Source: Elaborated by the authors.

Table 5. Interval between successive calls.

Day	Interval (hours)	Day	Interval (hours)
1	0.56	12	5.92
2	1.84	13	1.08
3	2.10	14	1.65
4	3.21	15	0.18
5	6.40	16	2.63
6	0.36	17	5.03
7	0.59	18	5.84
8	1.75	19	0.14
9	0.15	20	2.59
10	2.96	21	1.31
11	3.86	Average	2.39

Source: Elaborated by the authors.

Table 6. Arrival rate by atom λ_i (calls/hour).

Node	Calls	p_i	λ_i	Node	Calls	p_i	λ_i
01	278	1.56%	0.007	25	274	1.53%	0.006
02	278	1.56%	0.007	26	274	1.53%	0.006
03	236	1.32%	0.006	27	706	3.95%	0.017
04	396	2.22%	0.009	28	336	1.88%	0.008
05	166	0.93%	0.004	29	256	1.43%	0.006
06	406	2.27%	0.010	30	376	2.11%	0.009
07	166	0.93%	0.004	31	136	0.76%	0.003
08	256	1.43%	0.006	32	506	2.83%	0.012
09	334	1.87%	0.008	33	248	1.39%	0.006
10	334	1.87%	0.008	34	248	1.39%	0.006
11	334	1.87%	0.008	35	146	0.82%	0.003
12	334	1.87%	0.008	36	496	2.78%	0.012
13	286	1.60%	0.007	37	746	4.18%	0.017
14	166	0.93%	0.004	38	386	2.16%	0.009
15	406	2.27%	0.010	39	186	1.04%	0.004
16	316	1.77%	0.007	40	356	1.99%	0.008
17	526	2.94%	0.012	41	666	3.73%	0.016
18	396	2.22%	0.009	42	316	1.77%	0.007
19	1186	6.64%	0.028	43	261	1.46%	0.006
20	686	3.84%	0.016	44	261	1.46%	0.006
21	536	3.00%	0.013	45	1126	6.30%	0.026
22	493	2.76%	0.012	46	91	0.51%	0.002
23	493	2.76%	0.012	47	91	0.51%	0.002
24	274	1.53%	0.006	48	91	0.51%	0.002

Source: Elaborated by the authors.

Table 7. Average service time per server (minutes).

Server	Average service time	Standard deviation
1	59	41
2	95	67
3	81	65
4	156	94
5	18	14
6	86	77
7	43	30
8	114	68
9	29	23
ALS Average	77	54
BLS Average	75	53
Average	76	53

Source: Elaborated by the authors.

$$\overline{T} = \sum_{n=1}^{N}\sum_{i=1}^{I} f_{ni}^{[1]}\tau_{ni} + p_s\overline{T_Q} \quad (23)$$

$$\overline{T_i} = \left(\sum_{n=1}^{N} f_{ni}^{[1]}\tau_{ni} \Big/ \sum_{n=1}^{N} f_{ni}^{[1]}\right)(1-p_s) + \sum_{j=1}^{J}\left(\frac{\lambda_i}{\lambda}\right)t_{ji}p_s \quad (24)$$

In the previously equations, $f_{ni}^{[1]}$ is the fraction of the calls from atom *i* to server *n* that do not incur in awaiting time, such that $f_{ni}^{[1]} = (\lambda_i / \lambda)\sum_{B\in E_{ni}} p_B$, where λ_i / λ is the probability of a call from atom *i* to be is queued, and E_{ni} is the set of states in which a call from atom *i* is answer by server *n*. The traveling time of server *n* to atom *i* is given by:

$$\tau_{ni} = \sum_{m=1}^{I} g_{nm}t_{mi} \quad (25)$$

where $g_{nm} = 1$ when unit n is located at facility $j = i$, and 0, otherwise. On the other hand, p_s is de probability of saturation of the system, i.e., $p_s = p_Q + p_{11...1}$, such that $p_Q = 1 - (p_{00...0} + p_{00...10} + \ldots + p_{11...1})$ is the probability of queuing. Finally, $\overline{T_Q}$ is the average travel time for a call which is already awaiting, obtained by:

$$\overline{T_Q} = \sum_{j=1}^{J}\sum_{i=1}^{I} \frac{\lambda_j \lambda_i}{\lambda^2} t_{ji} \quad (26)$$

The application of HM to the data of the actual logistic arrangement leads to an average travel time of 13 minutes. Moreover, it was shown that only 42.4% of population is covered within the chosen "critical distance", i.e., 12 minutes.

4 Location model development

Two IP formulations from the literature review (FLEET and MALP) were used to locate the bases and position SAMU's ambulances in order to maximize covered demand. The following assumptions were considered:

i. Facilities opening and server's location: exactly one ambulance shall be placed at each open facility; although it is not a "hard" requirement, this approach is supported by the concept of decentralization, which tends to increase the covered population.

ii. Location constraints: primary and special ambulances can be located at any node, independently.

iii. Resources availability: limited number of servers given the current Health Department availability: seven primary ambulances and two advanced ones.

By setting $P^z = P^p + P^s$ in FLEET model, all the above premises are met. However, the use of backups is not considered, that is, the model does not take into account the redundancies that the system should have to avoid queuing and minimize service time in areas that have higher demands.

For this reason, the MALP was used as an alternative approach, once this model also meets the established premises. As seen, MALP deals with the stochastic nature of the problem by requiring a confidence level θ, which is guaranteed by using backups. The drawback in this case is possibility of considering only one type of server, thus requiring a simplifying assumption: advanced units shall be treated as basic units.

Given the size of the mathematical programs corresponding to the studied scenarios, they all could be optimized at low computational cost. For a subset of the best solutions found with the IP models, the performance indicators were evaluate by the HM. Obviously, such approach consists only of the analysis of some specific scenarios, and the use of Stochastic Optimization and Robust Optimization tend to offer better solutions. However, this bi-level strategy was able to identify some aspects that are likely to be part of an optimum stochastic solution. More precisely, facilities that have great chance to be opened, as well as the types of ambulances allocated to them.

5 Computational experiments

All tests were run on a Dell Inspiron 14R 3350 notebook with Intel Core™ i5 processor, under operating system Windows 7 Ultimate 64bit, powered with 6GB of RAM. AIMMS 3.13 was employed for coding the two IP models and the optimization of the mathematical programs was done through CPLEX 12.6. The best 200 solutions found by CPLEX for each program were exported to Microsoft Excel 2010, where a VBA code was used to implement the equations of the HM, thereby generating its respective performance indicators.

At first, FLEET was applied with pre-defined existing servers, that is, considering two advanced servers ($P^s = 2$) and seven basic servers ($P^p = 7$), and nine facilities ($P^z = 9$). By applying the HM, the best solution covered 66.3% of the population and resulted in an average time of 11.4 minutes per trip. The vehicles were positioned as follows: the advanced support ambulances at nodes 01 and 34, while the basic support at nodes 06, 12, 19, 20, 33, 41, and 46, one per location. Note that only the medical centers of Campos Elíseos and Saracuruna (nodes 06 and 41, respectively) which operate facilities at the time of this research are also locations provided by FLEET. This indicates how different the current arrangement is from that proposed by this approach, confirming the importance of using computational methods for solving this problem.

In turn, PLMD was solved considering $P = 9$ servers and no distinction among basic and advanced ambulances, in order to keep the homogeneity proposed by the model. For a confidence level θ = 93%, which corresponds to a coverage by at least $b = 5$ servers, there was obtained a solution in which 43.8% of the population is covered, within the response time defined by HM. This result can be explained by the higher level of reliability required, which tends to concentrate servers in more populous regions. The average travel time of the system was 14.8 minutes and ambulances were allocated to 01, 02, 09, 10, 11, 12, 30, 33, and 34. For θ = 88% of reliability, a coverage by at least four servers is required. The best solution evaluated by HM

predicts a coverage of 62.6% of the population with median time to onset of care 10.9 minutes, and the ambulances positioned at the nodes 01, 02, 06, 16, 19, 20, 23, 33, and 34. Finally, for θ = 80%, three servers are required, and HM produces a solution in which 72% of the population is covered up to 12 minutes, and servers were allocated to the nodes 02, 06, 09, 10, 16, 19, 20, 33, and 34. Interestingly, the ambulances were geographically grouped into two clusters along the network and the average travel time for system was 11 minutes. Table 8 summarizes the metrics discussed for each of the analyzed scenarios.

At this point, it is worth to point some considerations made by Chiyoshi et al. (2003) about the possibility of comparing coverage metrics calculated by different models. Metrics are not always comparable, especially in stochastic formulations, since the premises adopted are different incurring into practical implications. The same happens between IP models used in this paper, which differ in their underlying assumptions. To enable comparisons, the coverage metrics were computed by the HM, so all solutions could be evaluated by the same methodology.

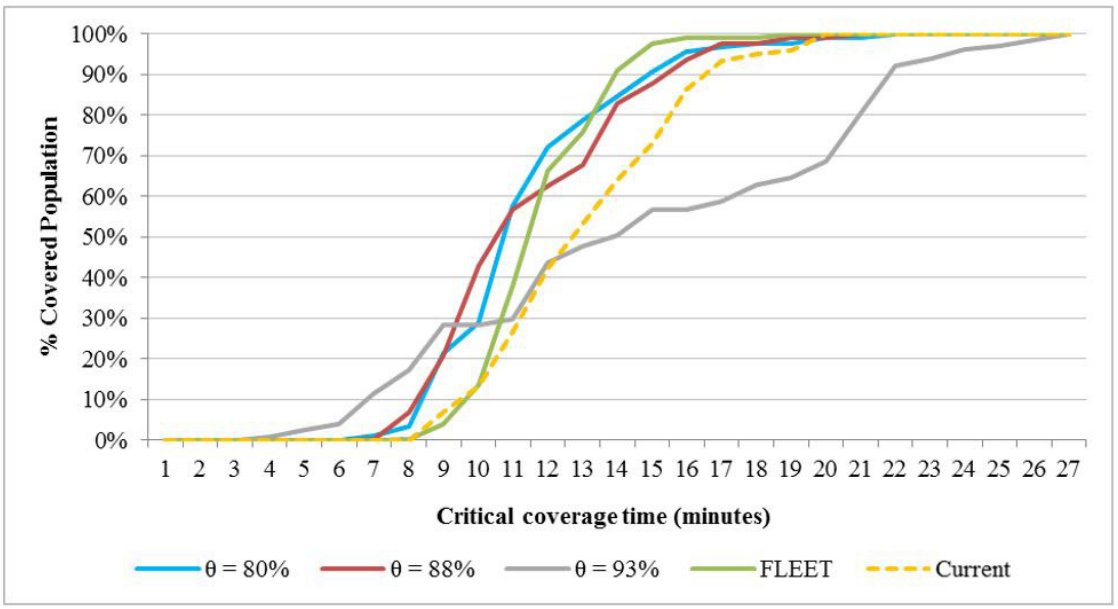
Despite the similarity among the mean times, the percentage of coverage varies considerably, because few changes in the location of the facilities may change the average. For example, when θ = 93%, MALP concentrates servers in order to reduce response time in specific regions of the network, which, in turn, increases the travel time to other network nodes, leaving them uncovered. However, the average time remains balanced. The same is true for system's configuration by the time of this research, whose average time value is close to those ones obtained by the optimization models. But it covers a significantly smaller population.

Graphic 1 shows the percentage of covered population according to response time. It is important to note that the solutions proposed by MALP provide smaller response times than the existing scenario at the time of this research and the FLEET model. Furthermore, for θ = 80% and θ = 88%, the results are always better than those of the current scenario.

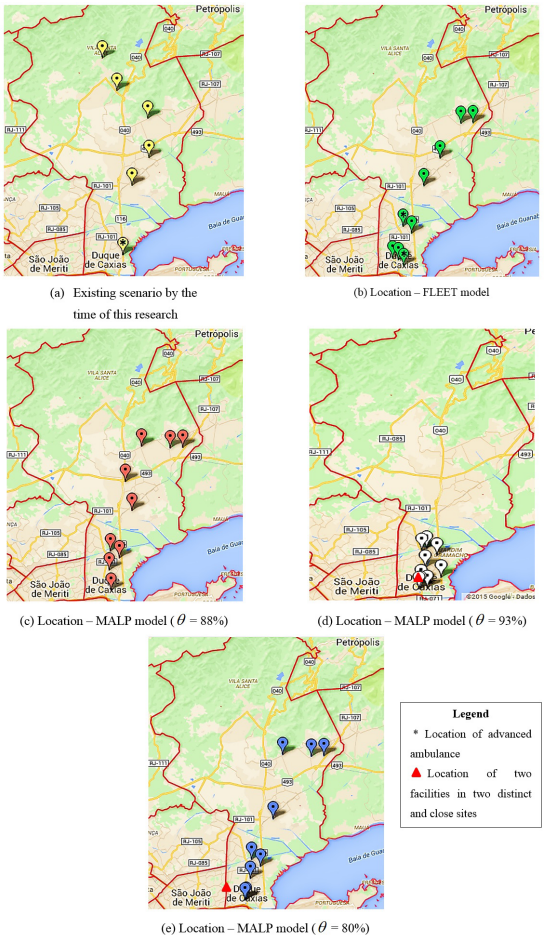
Table 8. Results comparing obtained by the model Hypercube.

Case	Covered demand (%)	Average travel time (min)
Current	42.4	13.0
FLEET	66.3	11.4
PLMD (θ = 93%)	43.8	14.8
PLMD (θ = 88%)	62.6	10.9
PLMD (θ = 80%)	72.0	11.0

Source: Elaborated by the authors.



Graphic 1. Covered population according to critical time. Source: Elaborated by the authors.



(a) Existing scenario by the time of this research

(b) Location – FLEET model

(c) Location – MALP model (θ = 88%)

(d) Location – MALP model (θ = 93%)

(e) Location – MALP model (θ = 80%)

Figure 2. Geographical distribution of servers. Source: Elaborated by the authors.

Their corresponding critical times for covering are almost the same, suggesting that the location has an important effect as the number of servers used as backups. It is also seen that for an upper critical time of 14 minutes, FLEET covers a percentage of the population bigger than than those covered by MALP. However, there must be paid attention to the fact that the first does not consider the effects of backup, which tends to reduce the chance of answering a call.

Figure 2 shows a comparison of the spatial distribution of ambulances in which: 2a shows the current logistic arrangement; 2b illustrates the solution obtained by applying the FLEET model; 2c, 2d and 2e illustrate the solutions obtained by MALP, for θ = 88%, θ = 93% and θ = 80%, respectively. In cases 2a and 2b, the marker with an asterisk ("*") indicates advanced ambulances. In the remaining cases there are no marks once they illustrate solutions of MALP, whose assumptions does not make distinction among servers.

While in existing configuration at the time of this research there were facilities only at six nodes, because of the concentration of four servers in the same node, the other models distribute ambulances along the network, according to demand, thence ensuring that more users are covered. Another important remark is the concentration of ambulances in the south, as shown in Figures 2b and 2d. This is the Duque de Caxias's downtown, its most populous region, from where the largest number of calls emerged. Finally, it is important to mention that Figures 2d and 2e use exactly nine facilities, but due to the limitations of map scale two geographically close locations were overlapping. Nevertheless, the scale was kept to ease comparisons among the images, and the overlapping facilities were highlighted with the mark "▲".

Moreover, one can also notice how MALP varies the geographic distribution of the ambulances as the confidence level θ increases. At Figures 2c and 2e, for θ = 88% and θ = 80%, respectively, we can see a trend forming two clusters, one in the south and another one in the northeast. At Figure 2d, it is seen that the highest level of reliability required (θ = 93%) yields a single cluster at south. Most likely, the availability of only nine ambulances does not stimulate the model to create other clusters, since there would be no gains in terms of coverage because the remaining nodes would not have five or more remote servers to answer their calls within 12 minutes.

6 Conclusion

In this work Optimization techniques and Queuing Theory were combined in order to analyze the SAMU's ambulance positioning in Duque de Caxias/RJ. It led to a significant increase in the coverage of the population and also provided a smaller response time for the same number of existing servers. The proposed solutions are significantly different from those used in existing configuration by the time of this research, reiterating the importance of mathematical-computational methods in location studies.

The solutions obtained by the IP models were used in HM aiming to evaluate the dispatch of ambulances and the system performance under congestion. The analyses indicate that MALP provides a better modeling of the stochastic behavior of the problem, thus leading to solutions with higher service level for the same values of service time. The results suggest a tendency to equilibrium between dispersion of servers and clustering in the network as an attempt to maximize coverage, while increases the chance of service by using backups.

It advisable that in future studies make a more extensive and accurate data collection, since this was one of the main difficulties found out during our investigation. Another important direction for further researches is the use of stochastic models, given the development of this area in recent years, particularly the Stochastic and Robust Optimization techniques. It would also be interesting to compare the quality and complexity of obtaining solutions for these models against a bi-level approach like the ones used in this paper.

References

Ball, M. O., & Lin, F. L. (1993). A reliability model applied to emergency service vehicle location. *Operations Research*, 41(1), 18-36. http://dx.doi.org/10.1287/opre.41.1.18.

Batta, R., Dolan, J. M., & Krishnamurthy, N. P. (1989). The maximal expected covering location problem: revisited. *Transportation Science*, 23(4), 277-287. http://dx.doi.org/10.1287/trsc.23.4.277.

Bertelli, A., Bueno, M. R., & Sousa, R. M. C. (1999). Estudo preliminar das relações entre duração da parada cardiorrespiratória e suas consequências nas vítimas de trauma. *Revista da Escola de Enfermagem da U S P.*, 33(2), 130-141. http://dx.doi.org/10.1590/S0080-62341999000200004.

Brandeau, M., & Larson, R. C. (1986). Extending and applying the hypercube queueing model to deploy ambulances in Boston. In A. J. Swersey & E. J. Ingnall (Eds.), *Delivery of urban services* (TIMS Studies in the Management Science, Vol. 22, pp. 121-153). London: Elsevier.

Brasil. Ministério da Saúde. (2013). *Portaria GM/MS n.º 2.657, 16 de dezembro de 2004. Estabelece as atribuições das centrais de regulação médica de urgências e o dimensionamento técnico para a estruturação e operacionalização das Centrais Samu - 192*. Brasília, DF: Diário Oficial da República Federativa do Brasil. Recuperado em 20 de agosto de 2013, de: http://bvsms.saude.gov.br/bvs/saudelegis/gm/2011/prt2026_24_08_2011.html.

Brotcorne, L., Laporte, G., & Semet, F. (2003). Ambulance location and relocation models. *European Journal of Operational Research*, 147(3), 451-463. http://dx.doi.org/10.1016/S0377-2217(02)00364-8.

Chiyoshi, F., Galvão, R. D., & Morabito, R. (2000). O uso do modelo Hipercubo na solução de problemas de localização probabilísticos. *Gestão & Produção*, 7(2), 146-174. http://dx.doi.org/10.1590/S0104-530X2000000200005.

Chiyoshi, F., Galvão, R. D., & Morabito, R. (2003). A note on solutions to the maximal expected covering location problem. *Computers & Operations Research*, 30(1), 87-96. http://dx.doi.org/10.1016/S0305-0548(01)00083-1.

Church, R. L., & Revelle, C. S. (1974). The maximal covering location problem. *Papers / Regional Science Association. Regional Science Association. Meeting*, 32(1), 101-118. http://dx.doi.org/10.1007/BF01942293.

Galvão, R. D., Chiyoshi, F., & Morabito, R. (2003a). Towards unified formulations and extensions of two classical probabilistic location models. *Computers & Operations Research*, 32(1), 15-33. http://dx.doi.org/10.1016/S0305-0548(03)00200-4.

Galvão, R., Chiyoshi, F., Espejo, L., & Rivas, M. A. (2003b). Solução do problema de localização de máxima disponibilidade utilizando o modelo Hipercubo. *Pesquisa Operacional*, 23(1), 61-78. http://dx.doi.org/10.1590/S0101-74382003000100006.

Gendreau, M., Laporte, G., & Semet, F. (1997). Solving an ambulance location model by tabu search. *Location Science*, 5(2), 75-88. http://dx.doi.org/10.1016/S0966-8349(97)00015-6.

Gendreau, M., Laporte, G., & Semet, F. (2001). A dynamic model and parallel Tabu search heuristic for real-time ambulance relocation. *Parallel Computing*, 27(12), 1641-1653. http://dx.doi.org/10.1016/S0167-8191(01)00103-X.

Ghiani, G., Laporte, G., & Musmanno, R. (2004). *Introduction to logistics systems planning and control*. West Sussex: John Wiley & Sons Ltd.

Iannoni, A. P., Morabito, R., & Saydam, C. (2009). An optimization approach for ambulance location and the districting of the response segments on highways. *European Journal of Operational Research*, 195(2), 528-542. http://dx.doi.org/10.1016/j.ejor.2008.02.003.

Instituto Brasileiro de Geografia e Estatística – IBGE. (2014). *Censo demográfico 2010*. Rio de Janeiro: IBGE. Recuperado em 26 de março de 2014, de http://cidades.ibge.gov.br/xtras/perfil.php?codmun=330170

Larson, R. C. (1974). A hypercube queuing model for facility location and redistricting in urban emergency services. *Computers & Operations Research*, 1(1), 67-95. http://dx.doi.org/10.1016/0305-0548(74)90076-8.

Larson, R. C. (1975). Approximating the performance of urban emergency service systems. *Operations Research*, 23(5), 845-868. http://dx.doi.org/10.1287/opre.23.5.845.

Revelle, C. S., & Hogan, K. (1989). The maximum availability location problem. *Transportation Science*, 23(3), 192-200. http://dx.doi.org/10.1287/trsc.23.3.192.

Schilling, D. A., Elzinga, D. J., Cohon, J., Church, R. L., & Revelle, C. S. (1979). The TEAM/FLEET models for simultaneous facility and equipment sitting. *Transportation Science*, 13(2), 163-175. http://dx.doi.org/10.1287/trsc.13.2.163.

Schmid, V., & Doerner, K. F. (2010). Ambulance location and relocation problems with time-dependent travel times. *European Journal of Operational Research*,

207(3), 1293-1303. PMid:21151327. http://dx.doi.org/10.1016/j.ejor.2010.06.033.

Souza, R., Morabito, R., Chiyoshi, F., & Iannoni, A. (2013). Análise da configuração de SAMU utilizando múltiplas alternativas de localização de ambulâncias. *Gestão & Produção*, 20(2), 287-302. http://dx.doi.org/10.1590/S0104-530X2013000200004.

Taha, H. A. (2008). *Pesquisa operacional* (8. ed.). São Paulo: Pearson Prentice Hall. 326 p.

Takeda, R., Widmer, J., & Morabito, R. (2004). Aplicação do modelo Hipercubo de filas para avaliar a descentralização de ambulâncias em um sistema urbano de atendimento médico de urgência. *Pesquisa Operacional*, 24(1), 39-71. http://dx.doi.org/10.1590/S0101-74382004000100004.

Toregas, C. R., Swain, R., Revelle, C. S., & Bergman, L. (1971). The location of emergency service facilities. *Operations Research*, 19(6), 1363-1373. http://dx.doi.org/10.1287/opre.19.6.1363.