Gest. Prod., São Carlos, v. 24, n. 4, p. 814-828, 2017
http://dx.doi.org/10.1590/0104-530X2452-16

Application of the hypercube model with queue priorities and more than one preferential server: a case study on a SAMU

Aplicação do modelo hipercubo com prioridade na fila com mais de um servidor preferencial sem considerar a hipótese de backup parcial: estudo de caso em um SAMU

Caio Vitor Beojone[1]
Regiane Máximo de Souza[1]

Abstract: The study of EMS aims to find ways to provide effective health services and improve the quality of life of the population while respecting the limitations of available resources. In this context, this paper aims to show the potential of application of the hypercube queueing model using queue priorities with more than one preferential server without using partial backup on SAMU, where the workload is relatively low. To do so, were done some experiments with the hypercube queueing model and future scenario prospection by a case study on the SAMU system from Bauru, Brazil. It was evaluated the impacts of demand increase over the system and the acquisition of a new ambulance was evaluated considering the best options to locate it. Main results show that a 50% demand increase can double mean response times. In contrast, minor increases have a smaller impact over the system, as observed on 5.71% and 13.57% demand increases, where the mean response times raised 5% and 16% respectively. The acquisition of a new ambulance was evaluated in terms of mean response times also. The best location had a 3% lower mean response time, on average.

Keywords: Emergency Medical Systems; Queueing theory; Hypercube model; OR in health care; SAMU.

Resumo: *O estudo de Sistemas de Atendimento Emergencial – SAE visa encontrar meios de fornecer serviços de saúde efetivos e melhorar a qualidade de vida da população respeitando as limitações de recursos disponíveis. Nesse contexto, o objetivo do presente trabalho foi mostrar o potencial de aplicação do modelo hipercubo com prioridade na fila com mais de um servidor preferencial sem considerar a hipótese de backup parcial em Sistemas de Atendimento Móvel de Urgência – SAMU em que o nível de utilização do sistema é relativamente baixo. Para isso foram realizados alguns experimentos do modelo hipercubo com prioridade na fila sem backup parcial e prospecção de cenários futuros por meio de um estudo de caso no SAMU da cidade de Bauru, SP. Foram avaliados os impactos do aumento na demanda sobre o sistema e o quanto e como (aonde localizar?) a aquisição de uma nova ambulância pode melhorar as medidas de desempenho do sistema. Os principais resultados mostram que um aumento de 50% na demanda pode dobrar o tempo de resposta dessas ambulâncias, por outro lado, aumentos mais discretos têm um impacto pequeno sobre o sistema, como pode ser visto nos aumentos de 5,71% e 13,57%, nos quais o acréscimo nos tempos de resposta foram de 5% e 16%, respectivamente. A aquisição de uma nova ambulância foi avaliada em termos das medidas de desempenho e os melhores resultados em todos os cenários se deu quando ela estava presente no átomo Boulevard, obtendo um tempo médio de resposta 3% inferior às demais localidades, em média.*

Palavras-chave: *Sistemas Médicos Emergenciais; Teoria das filas; Modelo hipercubo; PO em saúde; SAMU.*

1 Introduction

Around 140.000 traffic accidents happened in Brazil in the year of 2014 (DNIT, 2015). Even though it is a reduction when compared to years before, Brazil still is the third country with the highest number of traffic deaths in the world (OMS, 2016). That turns the pre-hospital service (one of the least expensive costs on traffic accidents) of great importance on the reduction in the number of deaths and severe injuries caused by traffic accidents (Lopes & Fernandes, 1999).

The pre-hospital service, as shown, is part of an Emergency Medical System (EMS). These systems are studied since the decade of 1950, using Operational Research, and have shown great potential for improvement on firefighters, police,

[1] Departamento de Engenharia de Produção, Universidade Estadual Paulista "Júlio de Mesquita Filho" – UNESP, Av. Eng. Luiz Edmundo C. Coube, 14-01, Vargem Limpa, CEP 17033-360, Bauru, SP, Brazil, e-mail: beojone@hotmail.com; regiane@feb.unesp.br

Received June 13, 2016 - Accepted Nov. 3, 2016

Financial support: CAPES and FAPESP.

Emergency Mobile Care Service (SAMU – "Serviço de Atendimento Móvel de Urgência"), and others. The main idea is to give effective health services and to improve the quality of life for the population considering the existent trade-off between the service level and the resource shortage (Simpson & Hancock, 2009; Souza, 2010).

SAMU is a public EMS created by Brazilian federal government from a bilateral agreement between Brazil and France. The system is based on the French model, operating for more than 30 years (Takeda et al., 2004). It works 24/7; the crews have physicians, nurses and nursing assistants. One can request the service using the telephone number 192. Requests are labeled according to their location and urgency level. An ambulance answers the request in the street, workplaces, or home (Ghussn & Souza, 2016).

Systems like SAMU are essentially characterized by uncertainties, mainly by availability, location, service time, demand through a neighborhood and response time to users. Therefore, EMS are a great challenge to health systems. Independently of the kind of urgency involved, only with a precise management, it is possible to offer good services (Souza, 2010).

The hypercube queueing model is a descriptive model based on Queueing Theory, which can calculate relevant performance measures for an EMS. We can divide the measures into two groups: external (from the user point of view) as mean response times, mean travel time, fraction of requests answered inside a time limit; and internal (from the system manager point of view) as the workload, dispatch frequencies, etc. (Larson & Odoni, 2007).

The hypercube queueing model was used on several papers in Brazil, as Chiyoshi et al. (2000), Iannoni et al. (2009, 2015), Iannoni & Morabito (2006, 2008), Souza et al. (2013, 2014, 2015), Takeda et al. (2004, 2007), Rodrigues (2014). In other parts of the globe, many other papers emerge as Chelst & Barlach (1981), Brandeau & Larson (1986), Burwell et al. (1993), Sacks & Grief (1994), Swersey (1994) and Larson & Odoni (2007). On every study, the hypercube has shown to be efficient and precise.

In this context, this paper aims to show the potential of application of the hypercube queueing model using queue priorities considering more than one preferential server without using partial backup on SAMU, where the workload is relatively low. To do so, we did some experiments with the hypercube queueing model and future scenario prospection by a case study on the SAMU system from Bauru, Brazil.

Section 2 shows a description of SAMU system and its service process. Section 3 comes with an explanation about the hypercube queueing model used along the paper and its particularities. Section 4 shows the hypotheses tests for the hypercube model application. Section 5 describes the results obtained, validating the model and the impacts of alternative scenarios. Finally, Section 6 shows final considerations and further research perspectives.

2 The SAMU system in Bauru

According to JCNet (2010), in 2010, the SAMU system in Bauru (SAMU/Bauru), Brazil, was responsible for serving 17 municipalities near Bauru through a partnership between them. The service headquarters is located on Bauru and is responsible for receiving the requests (calls) and distribute them according to their urgency level and location. There were around ninety professionals working on this operation, there were 35 drivers and 32 assistants, among them. The operation had 25 ambulances serving the 17 municipalities. The region under their responsibility has a population of 800,000 inhabitants, 400,000 thousand of them are in Bauru.

Data collection used reports from SAMU/Bauru for the years of 2012 and 2013. The collected data refers to ten days from September of each year. A summary of the data collection and the hypothesis verifications are in Section 4.

Figure 1 describes the service process of a request in SAMU. The process starts with a call, usually done by the telephone number 192. If there is an idle ambulance, the crew is sent to set up the ambulance and then it leaves the base as soon as the request is labeled. Then the ambulance starts travelling to the request location. We call the time interval between the request receiving and the arrival at the request location as response time. Now starts the service itself for the user, which may vary depending on the appropriate care. The patient is taken to a destination, which may be a hospital or a health center. After leaving the patient on its destination, the ambulance returns to the base, finishing the job (Schmid, 2012).

The SAMU service also includes the classification of requests (labelling) according to their urgency level. The labels receive colors to distinguish them: red (most serious and life risk), yellow, green and blue (least serious). A physician labels the requests according to the information given by the requester at the time of the call. The dispatch policy uses these labels to decide which ambulance will answer a request. ASV (Advanced Service Vehicles) can only answer to life risk requests (red label), and they are also the preferential servers to these requests. We call this policy as Partial Backup, because ASV cannot answer least urgent requests (yellow, green and blue labels). Particularly, the SAMU/Bauru has two ASV, differently from the study of Souza et al. (2014), which had only one on the system.

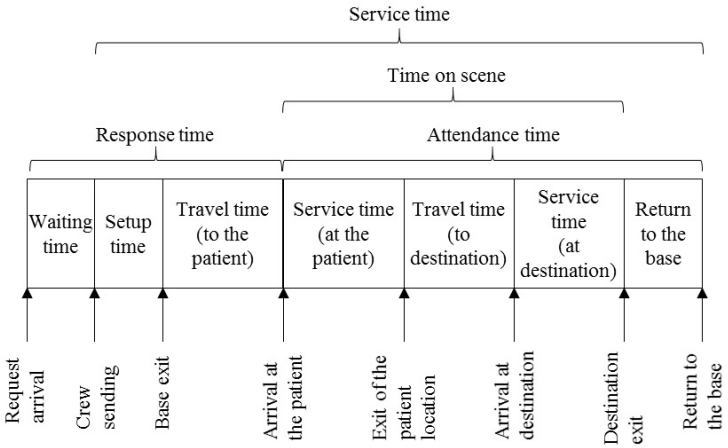

Figure 1. Time line with the events of an ambulance service.

3 Hypercube queueing model

Developed by Larson (1974), the hypercube queueing model is a descriptive model based on spatially distributed queueing systems. The idea is to expand the states of an *M/M/m* system to represent *m* servers individually, on a system where servers travel to the users. It allows working with more complex dispatch policies. In order to find a solution we need to prepare and solve a set of steady-state equations, the results are the probabilities of occurrence for each state. From the model solution, we can calculate several performance measures for the system as workloads, dispatch frequencies, mean travel times, mean waiting times, and others (Souza, 2010).

To use it on different approaches, we have to fit the classic model proposed by Larson (1974) in a way to turn the model closer to the reality of the analyzed system. We call these modifications as model extensions. Among the extensions, we can cite some, present on SAMU/Bauru, as priorities on queue and dispatch randomness.

The use of the hypercube model requires many parameters. The use of model extensions can change some of these parameters. Table 1 shows the notations for the parameters. Concepts about subatoms and priorities are shown on Section 3.3.

On the hypercube model, the state space indicates the availability of each server individually, as mentioned before. Consider a system with $m = 3$ servers, there are $2^3 = 8$ possible states on the system: {000}, {001}, {010}… {111}. Numbers 0 and 1 indicates whether a server is idle or on duty, respectively. A cube can represent the state space for three servers. In a case with more than three servers, we have a hypercube.

Figure 2 illustrates the state space for a system with three servers. It is possible to work with a system that allows or does not allow formation of waiting lines on the hypercube model. On the classic model, arriving calls wait in a line where the first server to become out of duty servers the users following a FCFS (First Come First Served) discipline. States S4, S5, S6 … on Figure 2 represent states with 1, 2, 3 … users on the waiting line of the system, respectively.

Larson & Odoni (2007) shows nine hypothesis that need to be satisfied to the application of the hypercube model on its classic shape.

1. Existence of geographical atoms: the region under service of the system must be divided into N_A geographical atoms, wherein each atom corresponds to an independent call source and has a dispatch policy.
2. Arrival process according to a Poisson: users of each atom request service by a Poisson process, wherein calls are independent between each other. Besides this, we must know the arrival rates, λ_j, from each atom.
3. Travel times of servers: we must know or estimate the travel times, τ_{ij}, of each server *i* to each atom *j*.
4. Servers: there are *N* servers spatially distributed over the system, wherein each one can travel and answer all atoms.
5. Location of servers: we must know the location of all servers (at least statistically). In other words, servers can move by atoms or stand still on one of them.
6. Simple dispatch: only one server answers a call. If there are no idle servers available, calls enter the waiting line or are considered loss of the system.

Table 1. Used notations.

Notation	Meaning	Measure unit
N_A	Number of atoms	Number
N	Number of servers	Number
τ_{ij}	Travel time from atom *i* to atom *j*	Minutes
$\tau_{ik,jh}$	Travel time from subatom *ik* to subatom *jh*	Minutes
λ_j	Arrival rate for atom *j* $\left(\lambda_j = \sum_k \lambda_{jk}\right)$	Calls/hour
λ_{jk}	Arrival rate for subatom *jk*	Calls/hour
λ_k	Arrival rate for priority *k*	Calls/hour
λ	System total arrival rate	Calls/hour
μ_i	Service rate for server *i*	Calls/hour
μ	System total service rate	Calls/hour
μ^{-1}	Service time	Minutes

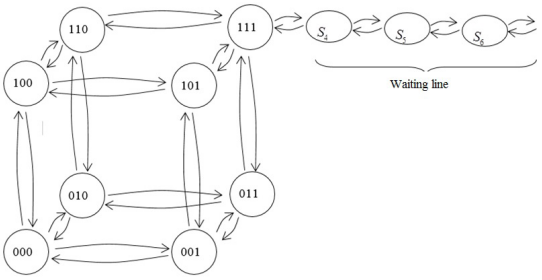

Figure 2. State space for the hypercube model with three servers.

7. Dispatch policy: there is a preference dispatch list (matrix) for all atoms, i. e., they must follow an order to send servers and it must be clear. For example, a system with four servers and three atoms, when a call from atom 3 arrives, we must follow the preference list (1, 3, 4, 2), i. e., server 1 is the first one sent, if the server is busy, send server 3 and so on until server 2.
8. Service time: the service time of a server includes the setup time, travel time, time on-scene until the return to the base. The service time follows an exponential distribution; and
9. Service time dependence on travel time: the variation of the travel time must be considered to be of the second order compared to the variations on the on-scene time and the setup time. It does not mean that we ignore travel times in computing mean service times; though the mean service time (μ^{-1}) calibration is equal to the sum of the mean travel time for the server, and the mean attendance time.

The extensions for the classic model change these hypotheses according to the necessity of the studied system. Examples of applications of the model on its original shape can be easily found on the literature for a better comprehension of its operation.

Now we present an application example of the hypercube classic model found in Chiyoshi et al. (2000). Consider a system, it do not accept waiting lines, three atoms form the region and they are served according to the dispatch preference list from Table 2.

The steady state equations show the behavior for the model in equilibrium (Equations 1 to 8). Note that, due to the dispatch preference list, the transition from state {100} to the state {110} has a $\lambda_1 + \lambda_2$ rate. This occurs because the preferential server for atom 1, server 1, is on duty and the second option is server 2, summed to the arrival rate from atom 2, where server 2 is the preferential server. We can solve the equations of the system as a determined homogeneous linear equation system if one of the steady state equations is substituted by the equation

$\sum_{B \in D} P_B = 1$ that shows that that sum of the probabilities of the system are equal to 1.

$$\lambda P_0 = \mu_1 P_{\{100\}} + \mu_2 P_{\{010\}} + \mu_3 P_{\{001\}} \tag{1}$$

$$(\lambda + \mu_1) P_{\{100\}} = \mu_2 P_{\{110\}} + \mu_3 P_{\{101\}} + \lambda_1 P_{\{000\}} \tag{2}$$

$$(\lambda + \mu_2) P_{\{010\}} = \mu_3 P_{\{011\}} + \mu_1 P_{\{110\}} + \lambda_2 P_{\{000\}} \tag{3}$$

$$(\lambda + \mu_3) P_{\{001\}} = \mu_2 P_{\{011\}} + \mu_1 P_{\{101\}} + \lambda_3 P_{\{000\}} \tag{4}$$

$$(\lambda + \mu_1 + \mu_2) P_{\{110\}} = \mu_3 P_{\{111\}} + (\lambda_1 + \lambda_2) P_{\{100\}} + \lambda_1 P_{\{010\}} \tag{5}$$

$$(\lambda + \mu_1 + \mu_3) P_{\{101\}} = \mu_2 P_{\{111\}} + \lambda_3 P_{\{100\}} + (\lambda_1 + \lambda_3) P_{\{001\}} \tag{6}$$

$$(\lambda + \mu_2 + \mu_3) P_{\{011\}} = \mu_1 P_{\{111\}} + (\lambda_2 + \lambda_3) P_{\{010\}} + \lambda_2 P_{\{001\}} \tag{7}$$

$$(\lambda + \mu) P_{\{111\}} = \lambda \left(P_{\{110\}} + P_{\{101\}} + P_{\{011\}} \right) \tag{8}$$

With the state probabilities in hand one can calculate many performance measures for the system.

Workload of a server is the mean time fraction that the server is busy. To calculate this important performance measure it is necessary to sum the probabilities that the server is busy, the state probabilities wherein the server is busy {1} plus the waiting line probabilities.

Another important performance measure is the dispatch frequencies. Its calculation includes dispatches without queue delay (*nq*) and dispatches with queue delay (*q*) as shown on Equation 9.

$$f_{ij} = f_{ij}^{(nq)} + f_{ij}^{(q)} = \frac{\lambda_j}{\lambda} \sum_{B \in E_{ij}} P_B + \frac{\lambda_j}{\lambda} P_Q^{'} \frac{\mu_i}{\mu} \quad (9)$$

where λ_j / λ is the arrival fraction corresponding to atom *j*.

$\sum_{B \in E_{nj}} P_B$ is the sum of the state probabilities that the server *n* is idle and is the next on the preference list to answer atom *j*. $P_Q^{'}$ is the system saturation probability, the sum of queue states plus the probability that all servers are busy. Finally, μ_i / μ is the probability that server I will be the first to finish a job when all servers are busy.

Equation 10 gives the mean time for server *i* travel to atom *j*, when available.

$$t_{ij} = \sum_{k=1}^{N_A} l_{ik} \tau_{kj} \quad (10)$$

Equation 11 calculates the mean queued request travel time.

$$\overline{T}_Q \equiv \sum_{i=1}^{N_A} \sum_{j=1}^{N_A} \frac{\lambda_i \lambda_j}{\lambda^2} \tau_{ij} \quad (11)$$

Equation 12 can calculate the mean travel time for the system.

$$\overline{T} = \sum_{i=1}^{N} \sum_{j=1}^{N_A} f_{ij}^{(nq)} t_{ij} + P_Q^{'} \overline{T}_Q \quad (12)$$

Equation 13 shows the mean travel time to atom *j*.

$$\overline{T}_j = \frac{\sum_{i=1}^{N} f_{ij}^{(nq)} t_{ij}}{\sum_{i=1}^{N} f_{ij}^{(nq)}} \left(1 - P_Q^{'}\right) + \sum_{k=1}^{N_A} \left(\frac{\lambda_k}{\lambda}\right) \tau_{kj} P_Q^{'} \quad (13)$$

Equation 14 calculates the mean travel time for server *i*.

$$\overline{TU}_i = \frac{\sum_{j=1}^{N_A} f_{ij}^{(nq)} t_{ij} + \frac{\mu_i}{\mu} \overline{T}_Q P_Q^{'}}{\sum_{i=1}^{N} f_{ij}^{(nq)} + \frac{\mu_i}{\mu} P_Q^{'}} \quad (14)$$

To calculate waiting times, we can use Little's Formula, as shown on Equation 15 (Little, 2011).

$$W_Q = \frac{L_Q}{\lambda \left(1 - P_{perda}\right)} \quad (15)$$

3.1 Priority in queues

Systems that consider priorities in a queue define classifications for users in terms of their priority. For modelling purposes, geographical atoms are divided in layers so each layer represents a specific priority (Takeda et al., 2007). Figure 3 illustrates the layering process for atoms requests on a system with three different priorities: *a*, *b*, and *c*. These layers form nonphysical subatoms of the original atoms.

Priority means the order that the requests are answered in the queue following the urgency level of the request, i.e. its classification (Souza et al., 2015). Figure 4 shows an example shown in Souza et al. (2015) illustrating transitions between queue states with priority. The nodes represent the system states (situations), for example, state {*abb*} is the situation where we have a priority *a* (high priority) request and two priority *b* (average priority) requests on the waiting line. The arcs that unite states represent the transitions between states, for example, the system goes out of the situation {*ab*} and enters on the situation {*b*} when a server finishes its job and starts to serve the first queued request. Note that priority *a* has higher priority compared to priority *b* and so priority *b* over priority *c*. So, while there is a priority

Table 2. Dispatch preference list for the example.

Atom	Preference		
	1º	2º	3º
1	1	2	3
2	2	3	1
3	3	1	2

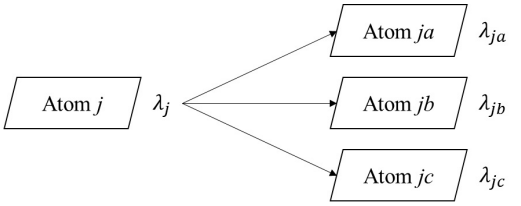

Figure 3. Layering process ilustration.

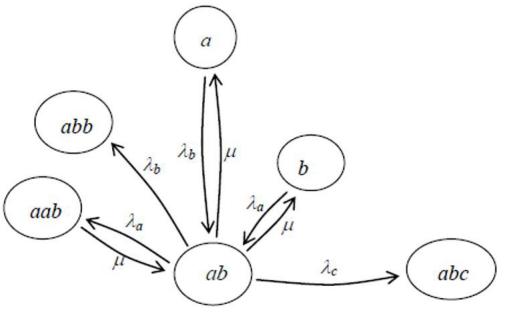

Figure 4. State transitions for queued requests with priority. Source: Souza et al. (2015, p. 276).

a request in a queue priorities *b* and *c* will not be served. Souza (2010) shows illustrative examples for this extension.

One can prepare steady state equations from these state transitions for the queue states with priorities. Equation 16 shows the equation for the state $\{ab\}$ shown on Figure 4.

$$(\lambda+\mu)P_{\{ab\}} = \lambda_a P_{\{b\}} + \lambda_b P_{\{a\}} + \mu P_{\{aab\}} \tag{16}$$

The main alteration on performance measures is the possibility to calculate mean waiting times for each priority (Souza et al., 2015). Considering n_r as the number of class *r* users, Equation 17 gives sum of probabilities that a state *S* has *j* users of priority *r* in a queue.

$$P(n_r = j) = \sum_{\forall S\, s.t.\ n(r,S)=j} P\{S\} \tag{17}$$

where S is the queue state and $n(r, S)$ is the number of class *r* users on *S*. Equation 18 enumerates the mean number of class r users in a queue, following the distribution from Equation 10.

$$L_{qr} = \sum_j jP(n_r = j) \tag{18}$$

So, the mean waiting time in a queue is obtained by Little's Formula again (Equation 19).

$$W_{qr} = L_{qr} / \lambda_r \tag{19}$$

3.2 Dispatch randomness

This extension tries to represent systems where a dispatch preference matrix is not well defined. Cases which once a server has the preference to answer a request, but another time other server may have the preference, and what decides it is not clear, it is a random choice.

There are at least two different ways to represent such system. Firstly, we can model and solve many models with a randomly created dispatch preference lists for each solution, as seen in Takeda et al. (2004). By the end of the solutions, a mean value is calculated to find the final probabilities for each state. On a different way, according to Chiyoshi et al. (2011), is to solve a single model where the transition rates are shared among idle servers. For example, on a transition between {000} to {001}, where all servers on the same location and there is no difference of their dispatch preference, the transition rate is equal to 1/3 of the total arrival rate. Equation 20 illustrates this case.

$$\lambda P_{\{100\}} = \frac{\lambda}{3} P_{\{000\}} + \mu_2 P_{\{110\}} + \mu_3 P_{\{101\}} \tag{20}$$

Larson (1974) and Batta et al. (1989) show a way that uses "correction factors" to choose which server answers a request. The idea is that the closest server answers the request and server location is statistically known. The probability that a specific server answers the request is proportional to the product of the probability that this server is available and the probability of the preferential server is busy. This process is not exactly random, as the "correction factors" really are deterministic.

4 Hypotheses verification for hypercube model application

4.1 Existence of N_A geographic atoms

In the year of 2013, the SAMU/Bauru separated its requests according to their place of origin, geographic atoms. Figure 5 contains the city map with the area division.

The SAMU/Bauru divides the requests according to their urgency levels: red (urgent requests with life risk), yellow (serious emergency), green (moderate emergency), blue (light emergency). Therefore, atoms were divided into layers, forming subatoms.

4.2 Arrival process

Using Kolmogorov Smirnov fit test for the time interval between arrivals, we verified that they follow and exponential distribution, confirming the second hypothesis of hypercube model. Figure 6 shows the test results considering a significance level of 0.05, where the hypothesis that the data follows an exponential distribution could not be rejected, observing the p-value of 0.172.

The calculations for each subatom arrival rate is done multiplying the proportion of requests from subatom (p_{jk}) and the total arrival rate of the system. Table 3 shows the identification, according to the atom and priority label, for each subatom and their respective arrival rate.

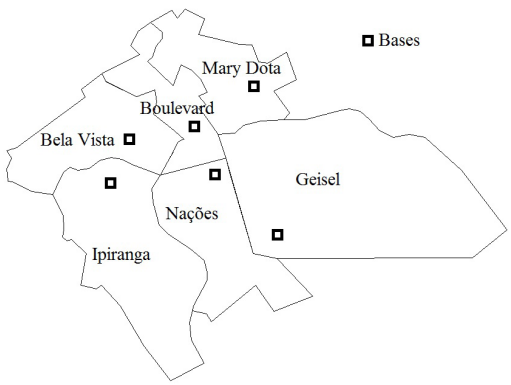

Figure 5. Bauru city map and its atoms and bases.

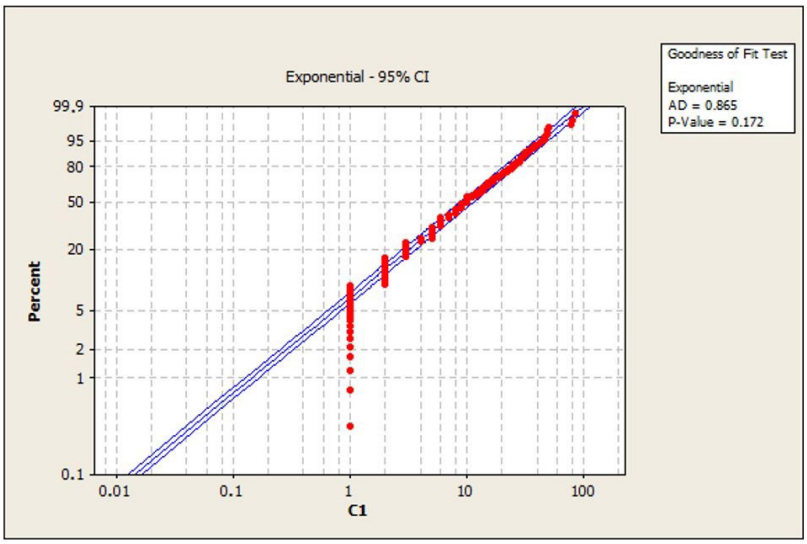

Figure 6. Fit test for exponential distribution.

Table 3. Atom and subatom identification and their arrival rates.

Subarea	Atom	Subatom	N° of requests	p_{jk}	λ_{jk} (requests/hour)
Geisel blue	1	1*d*	1	0.0050	0.0174
Geisel green	1	1*c*	12	0.0594	0.2090
Geisel yellow	1	1*b*	15	0.0743	0.2612
Geisel red	1	1*a*	5	0.0248	0.0871
Nações blue	2	2*d*	5	0.0248	0.0871
Nações green	2	2*c*	20	0.0990	0.3483
Nações yellow	2	2*b*	23	0.1139	0.4005
Nações red	2	2*a*	3	0.0149	0.0522
Ipiranga blue	3	3*d*	3	0.0149	0.0522
Ipiranga green	3	3*c*	11	0.0545	0.1916
Ipiranga yellow	3	3*b*	9	0.0446	0.1567
Ipiranga red	3	3*a*	4	0.0198	0.0697
Mary Dota green	4	4*c*	15	0.0743	0.2612
Mary Dota yellow	4	4*b*	7	0.0347	0.1219
Mary Dota red	4	4*a*	2	0.0099	0.0348
Bela Vista blue	5	5*d*	7	0.0347	0.1219
Bela Vista green	5	5*c*	18	0.0891	0.3134
Bela Vista yellow	5	5*b*	23	0.1139	0.4005
Bela Vista red	5	5*a*	6	0.0297	0.1045
Boulevard blue	6	6*d*	1	0.0050	0.0174
Boulevard green	6	6*c*	3	0.0149	0.0522
Boulevard yellow	6	6*b*	6	0.0297	0.1045
Boulevard red	6	6*a*	1	0.0050	0.0174
SAMU/Bauru			202	1.0000	3.5176

4.3 Travel times

Mean travel times between atoms were estimated using the data sample. In cases where there was not observations, we used the software Google Earth® to create an estimative. With the software, we could calculate the distance between atoms centroids, and supposing a 60km/h travel speed, we estimated mean travel times. Table 4 shows all travel times, an asterisk (*) indicates those that did not come from the sample, they were calculated using the explained method.

4.4 Servers

SAMU/Bauru has nine ambulances in Bauru. Wherein, two of them are ASV and seven are BSV (Basic Service Vehicles). Moreover, all ambulances

Table 4. Mean travel times between subatoms (minutes). The data indicated by an asterisk (*) are obtained by estimates means considering the distances of the centroids of the atoms and assuming an average speed of 60 km / h.

$\tau_{ik,jh}$	**1** ***a***	**1** ***b***	**1** ***c***	**1** ***d***	**2** ***a***	**2** ***b***	**2** ***c***	**2** ***d***	**3** ***a***	**3** ***b***	**3** ***c***	**3** ***d***	**4** ***a***	**4** ***b***	**4** ***c***	**4** ***d***	**5** ***a***	**5** ***b***	**5** ***c***	**5** ***d***	**6** ***a***	**6** ***b***	**6** ***c***	**6** ***d***
1 ***a***	10	10	10	10	13	13	13	13	14	14	14	14	14	14	14	14	14*	14*	14*	14*	11	11	11	11
1 ***b***	10	10	10	10	13	13	13	13	14	14	14	14	14	14	14	14	14*	14*	14*	14*	11	11	11	11
1 ***c***	10	10	10	10	13	13	13	13	14	14	14	14	14	14	14	14	14*	14*	14*	14*	11	11	11	11
1 ***d***	10	10	10	10	13	13	13	13	14	14	14	14	14	14	14	14	14*	14*	14*	14*	11	11	11	11
2 ***a***	13	13	13	13	8	8	8	8	11	11	11	11	29	29	29	29	9	9	9	9	6	6	6	6
2 ***b***	13	13	13	13	8	8	8	8	11	11	11	11	29	29	29	29	9	9	9	9	6	6	6	6
2 ***c***	13	13	13	13	8	8	8	8	11	11	11	11	29	29	29	29	9	9	9	9	6	6	6	6
2 ***d***	13	13	13	13	8	8	8	8	11	11	11	11	29	29	29	29	9	9	9	9	6	6	6	6
3 ***a***	14	14	14	14	11	11	11	11	8	8	8	8	18*	18*	18*	18*	15	15	15	15	13*	13*	13*	13*
3 ***b***	14	14	14	14	11	11	11	11	8	8	8	8	18*	18*	18*	18*	15	15	15	15	13*	13*	13*	13*
3 ***c***	14	14	14	14	11	11	11	11	8	8	8	8	18*	18*	18*	18*	15	15	15	15	13*	13*	13*	13*
3 ***d***	14	14	14	14	11	11	11	11	8	8	8	8	18*	18*	18*	18*	15	15	15	15	13*	13*	13*	13*
4 ***a***	14	14	14	14	29	29	29	29	18*	18*	18*	18*	8	8	8	8	5	5	5	5	11	11	11	11
4 ***b***	14	14	14	14	29	29	29	29	18*	18*	18*	18*	8	8	8	8	5	5	5	5	11	11	11	11
4 ***c***	14	14	14	14	29	29	29	29	18*	18*	18*	18*	8	8	8	8	5	5	5	5	11	11	11	11
4 ***d***	14	14	14	14	29	29	29	29	18*	18*	18*	18*	8	8	8	8	5	5	5	5	11	11	11	11
5 ***a***	14*	14*	14*	14*	9	9	9	9	15	15	15	15	5	5	5	5	8	8	8	8	7	7	7	7
5 ***b***	14*	14*	14*	14*	9	9	9	9	15	15	15	15	5	5	5	5	8	8	8	8	7	7	7	7
5 ***c***	14*	14*	14*	14*	9	9	9	9	15	15	15	15	5	5	5	5	8	8	8	8	7	7	7	7
5 ***d***	14*	14*	14*	14*	9	9	9	9	15	15	15	15	5	5	5	5	8	8	8	8	7	7	7	7
6 ***a***	11	11	11	11	6	6	6	6	13*	13*	13*	13*	11	11	11	11	7	7	7	7	7*	7*	7*	7*
6 ***b***	11	11	11	11	6	6	6	6	13*	13*	13*	13*	11	11	11	11	7	7	7	7	7*	7*	7*	7*
6 ***c***	11	11	11	11	6	6	6	6	13*	13*	13*	13*	11	11	11	11	7	7	7	7	7*	7*	7*	7*
6 ***d***	11	11	11	11	6	6	6	6	13*	13*	13*	13*	11	11	11	11	7	7	7	7	7*	7*	7*	7*

can travel between atoms, confirming the fourth model hypothesis.

4.5 Server location

Each server is located on a fixe base. Both ASV are located on SAMU/Bauru headquarters, on Geisel. While BSV are located on different bases, as follows: one in Geisel, one in Nações, one in Ipiranga, one in Mary Dota, two on Bela Vista and one in Boulevard.

4.6 Simple server dispatch

In the great majority, around 96% of all cases, requests where served by only one ambulance. Only in some cases, less than 4%, there was the dispatch of more than one vehicle to the scene. Moreover, ASV answer only red requests, which does not comply with the hypercube model hypothesis, but the workload over the system is low, allowing a relaxation of this hypothesis as seen in Takeda et al. (2004, 2007). The system allows the formation of waiting lines and there is no limit for its size, wherein requests are organized following their priority.

4.7 Dispatch preference

The server dispatch policy take some aspects in count, they are request location, urgency level and servers locations. The preferential server is the one located at the request atom, if this server is busy, a server from an adjacent atom is chosen. However, for red requests ASV are preferential servers regardless of the request location. These aspects cause dispatch randomness, i.e. it is not possible to write a fixed dispatch preference list (Burwell et al., 1993; Takeda et al., 2007).

4.8 Service times

Each server had its service time tested on an exponential curve fit test. We used Kolmogorov-Smirnov test on all servers and the hypothesis was rejected on a significance level of 0.05. Moreover, we tested if servers were homogeneous using an Analysis of variance (ANOVA) on a significance level of 0.05. Results show differences between their mean service times on all BSV, except those located on Bela Vista, so that servers are heterogeneous except ASV and Bela Vista's BSV.

Table 5 shows mean service rates for each server, obtained using their mean service times.

4.9 Service time dependence on travel time

Table 6 compares mean service times and the mean travel times. One can note that the mean travel times are relatively small, compared to mean service times, as it never represents more than 25%. This validates the hypothesis for the hypercube model application.

5 Model application and results analysis on SAMU/Bauru

The used model on SAMU/Bauru was the hypercube model, because the system is too complex to apply simpler queueing models than the hypercube model and almost no hypercube model hypotheses could be rejected considering the hypercube extensions of priority in a queue and dispatch randomness. The priority in a queue extension was chosen because SAMU/Bauru label requests according to their urgency level. The dispatch randomness extension was chosen because the system do not obey a fixed dispatch preference list.

It is worth mention that the dispatch randomness method used was the random creation of dispatch preference lists. Model was programmed on a Pascal interface. We ran the model 50 times before calculating the mean probabilities and, if necessary, we did a service time calibration considering a 0.1 requests/hour tolerance.

The first model (original), used on model validation, compared workloads and mean travel times. We did not compare waiting times because there was no data about screening times, which affects waiting times inside the sample.

Table 5. Mean service rates on SAMU/Bauru.

Ambulances		Mean service time (minutes)	μ_i (requests/hour)
1 -	GA1 (VSA)	56.8	1.0562
2 -	GA2 (VSA)	56.8	1.0562
3 -	GB (VSB)	47.2	1.2719
4 -	NÇ (VSB)	40.8	1.4720
5 -	IP (VSB)	42.9	1.3998
6 -	MD (VSB)	47.9	1.2517
7 -	BV1 (VSB)	49.0	1.2254
8 -	BV2 (VSB)	49.0	1.2254
9 -	BLV (VSB)	45.8	1.3089

Table 6. Relation between mean service times and mean travel times of servers.

Ambulances		Mena service time (minutes)	Mean travel times (minutes)	Proportion
1 -	GA1	56.8	13.6	0.2389
2 -	GA2	56.8	13.6	0.2389
3 -	GB	47.2	10.8	0.2281
4 -	NÇ	40.8	8.4	0.2056
5 -	IP	42.9	8.8	0.2047
6 -	MD	47.9	10.0	0.2079
7 -	BV1	49.0	9.0	0.1838
8 -	BV1	49.0	9.0	0.1838
9 -	BLV	45.8	7.3	0.1597

5.1 Original scenario *versus* sample

The results for the original model was compared to the sample data. Performance measures of both validates de used of the hypercube model on SAMU/Bauru.

Table 7 shows the workloads obtained on the original model and those obtained on the sample. It also compares their relative deviation. This measure had a good adherence to the sample with a mean deviation of 8%. One can note that ASV (GA1 and GA2) had a 7.4% increase compared to the sample, this could have happened because we did not restricted ASV service to red requests. However, this deviation did not compromise the model validation. The ambulance form Mary Dota (MD) had the highest error, 15%, which was considered acceptable, considering the mean results.

Table 8 shows subatoms mean travel times comparisons. The mean relative deviation was 9%. In the other hand, due to the small size of the sample for some subatoms (only 5 requests or less for 11 out of 24 subatoms) the data for labels *b*, *c* and *d* were aggregated. It diminished the deviations found, without losing precision, since they receive service from the same ambulances and have the same dispatch preference lists. The highest deviations, but considered acceptable, were from priority *a* subatoms, because of their small samples and low arrival rates (approximately 10% of requests) and they could not be aggregated with other priorities.

Lastly, we analyzed servers mean travel times. Table 9 shows the comparisons for these mean times between the model and the sample. Here, the mean deviation is lower than 6%, presenting the model good adherence. Sample size helps here, because all servers had at least 20 requests in the sample, some of them also had more than 50 (BV1 and BV2). The largest observed deviation (ambulance NÇ) was from the server with the smallest sample, with 21 requests.

Another number that shows the model adherence to the sample is the mean travel time of the system, which was 9.7 minutes, less than 1% deviation from the sample. Besides, there was no need to calibrate mean service times, a good information about the model adherence.

As results from workloads and mean travel times show a good adherence to the real system, one can consider the model as validated, turning possible to analyze alternative scenarios. Firstly, we studied the effects of demand increasing, based on SAMU/Bauru data from the years of 2012 and 2013. Then the inclusion of a new ambulance in order to mitigate middle and long-term demand increase was analyzed.

5.2 Demand increase

For demand increase scenarios, we forecasted some system trends for the years of 2014 (sample coming year). Three different methods were used to obtain the forecasts:

- Requests proportion (13.57%): with an increase on requests proportion on demand peak periods from 2012 and 2013, we suppose that this increase could be the same for coming year of 2014;
- Time series analysis (5.71%): using 2012 and 2013 requests summary, we tried to find forecasts to September 2014 using decomposition method, where September was chosen arbitrarily;
- Other forecasts (25% and 50%): arbitrary increase choices to represent extreme untypical situations.

Demand increase scenarios show the impact over the system caused by the increase of the mean number of arriving users on the system, turning it more congested. This makes mean waiting times to increase, the number of requests served by backups (non-preferential ambulances) to increase.

Figure 7 shows these impacts over mean response times of servers. It is important to note that a demand increase 50% can double the response time for these ambulances. On the other hand, modest demand increases have small impact over the system, for

example, 5.71% and 13.57% demand increases raise response times in 5% and 16% respectively.

The impact is visible over urgency levels too. Figure 8 shows different effects over each urgency level label of the system. Users with higher priorities have larger mean response times on initial scenarios because ASV serves them and ASV usually must cross the entire system to serve. However, the demand

Table 7. Model and sample workloads comparison.

Ambulance	Sample	Model	Relative deviation
GA1	0.1657	0.1779	7.4%
GA2	0.1657	0.1779	7.4%
GB	0.3800	0.3580	-5.8%
NÇ	0.3343	0.3690	10.4%
IP	0.2619	0.2884	10.1%
MD	0.3994	0.3388	-15.2%
BV1	0.3672	0.3506	-4.5%
BV2	0.3672	0.3506	-4.5%
BLV	0.3183	0.3485	9.5%

Table 8. Model and sample mean travel times (minutes) for subatoms comparison.

Subatoms	Sample	Model	Relative deviation
1 *a*	9.6	10.1	4.9%
1 *b*	10.7	11.2	4.7%
1 *c*	10.7	11.2	4.7%
1 *d*	10.7	11.2	4.7%
2 *a*	12.7	13.1	3.3%
2 *b*	8.6	8.5	-0.9%
2 *c*	8.6	8.5	-0.9%
2 *d*	8.6	8.5	-0.9%
3 *a*	13.3	13.6	2.7%
3 *b*	10.0	9.6	-4.2%
3 *c*	10.0	9.6	-4.2%
3 *d*	10.0	9.6	-4.2%
4 *a*	10.5	13.3	27.1%
4 *b*	8.5	9.9	17.5%
4 *c*	8.5	9.9	17.5%
4 *d*	8.5	9.9	17.5%
5 *a*	17.0	13.6	-20.2%
5 *b*	8.7	9.0	3.4%
5 *c*	8.7	9.0	3.4%
5 *d*	8.7	9.0	3.4%
6 *a*	14.0	10.9	-22.4%
6 *b*	9.6	8.1	-15.3%
6 *c*	9.6	8.1	-15.3%
6 *d*	9.6	8.1	-15.3%

Table 9. Model and sample mean travel times (minutes) for servers comparison.

Ambulance	Sample	Model	Relative deviation
GA1	13.6	12.7	-6.6%
GA2	13.6	12.7	-6.6%
GB	10.8	11.0	2.5%
NÇ	8.4	9.6	14.8%
IP	8.8	9.6	9.9%
MD	10.0	10.4	4.1%
BV1	9.0	8.9	-1.5%
BV2	9.0	8.9	-1.5%
BLV	7.3	7.0	-4.0%

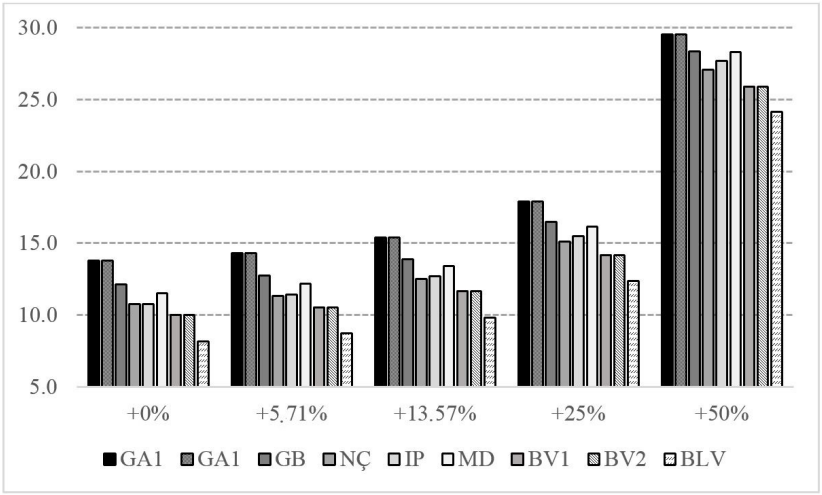

Figure 7. Demand increase impact over mean response times (minutes) of servers.

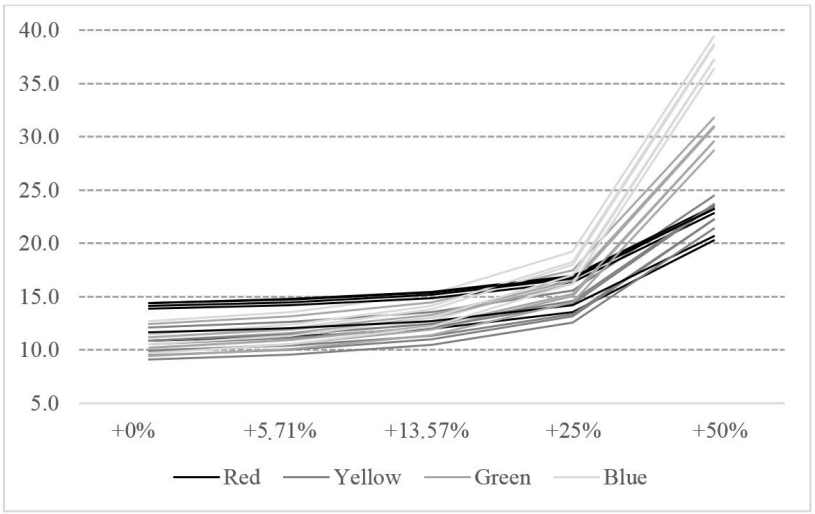

Figure 8. Demand increase impact over mean response times (minutes) for subatoms.

increases have minor impact over them. Users of smaller priorities, served by local BSV, have smaller response times (around 11 minutes on initial scenarios). Although, demand increases have major impact over them, taking them to a mean of 34 minutes for green and blue requests. This also shows the importance to represent different urgency levels.

5.3 New server inclusion

For each demand increase scenario, we analyzed the impacts over performance measures of including a new BSV. The ambulance was tested on all six existing bases (atoms). It used the same data (location and attendance time) as the ambulances already on those bases. The demand increase effects mitigation were evaluated using mean response time of the system.

Figure 9 compares the impact over mean response time caused by the inclusion of the new ambulance. Note that the model enables us to find the best location for the new ambulance and to quantify the effect over the performance measures. Thus, in the case the manager wants to be prepared for a period with a 25% demand increase, it is possible to maintain the same mean response time with a single new ambulance over atom Boulevard, mitigating demand increase effects.

The best location for the new ambulance was over Boulevard atom on all scenarios, with a 3% average lower mean response time than other locations. On the other hand, putting the ambulance over Mary Dota atom gives the least advantageous results until a 25% demand increase. On the most serious scenario, with 50% demand increase, the two best locations

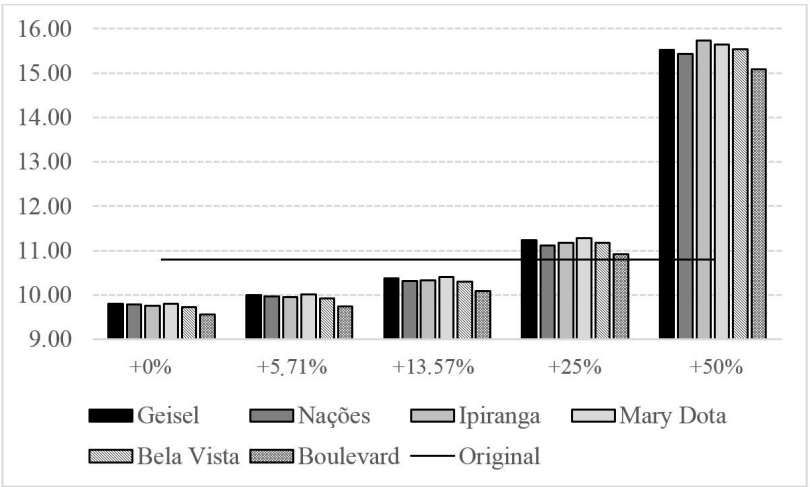

Figure 9. Mean response times (minutes) comparison with the inclusion of a new ambulance.

are Nações and Boulevard atoms, both are central regions of the city.

6 Final considerations

SAMU/Bauru is a system that serves a population using four different urgency levels; the service quality is deeply connected to a good resource management. Thus, this paper aims to show the potential of application of the hypercube queueing model using queue priorities with more than one preferential server without using partial backup on SAMU, where the workload is relatively low. To do so, we did some experiments with the hypercube queueing model with queue priority without partial backup and future scenario prospection by a case study on the SAMU system from Bauru, Brazil.

The deviations from the sample were considered acceptable for the model, enableing us to analyze alternative scenarios. These scenarios evaluated the impact of demand growth and the options to include a new ambulance. The model gives a analytical and detailed vision of the impacts caused by decisions made while managing a EMS. So that, it can improve the service to the population with better response times and appropriate ambulance dispatch. It is possible because the model enables evaluation and prospection of future situations, when properly calibrated to the analyzed system.

On alternative scenarios it was possible to find that a 50% demand increase can double mean response times. On the other hand, minor increases have a smaller impact over the system, as observed on 5.71% and 13.57% demand increases, where the mean response times raised 5% and 16% respectively.

Acquiring a new ambulance was evaluated in terms of performance measures and the best results on all scenarios occurred when this ambulance was over Boulevard atom, obtaining a 3% lower result than other locations, on average. On the other hand, putting the ambulance over Mary Dota atom gives the least advantageous results until a 25% demand increase. On the most serious scenario, with 50% demand increase, the two best locations are Nações and Boulevard atoms, both are central regions of the city.

For further research is proposed to use the model considering partial backup and queue, with restrictions for urgency levels as seen for the ASV on SAMU/Bauru, where they only serve red requests. Moreover, evaluate dynamic server positioning using time dependent models capable of finding the effects of system variations through the day.

Acknowledgements

The authors acknowledge CAPES and FAPESP for their research financial support. They also acknowledge the SAMU from Bauru for the availability of data and to the anonymous reviewer for the precious comments, which were extremely valuable for the evolution of this paper.

References

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

Brandeau, M., & Larson, R. (1986). Extending and applying the hypercube queueing model to deploy ambulances in Boston. In A. Swersey & E. Ingnall, *Delivery of urban services: studies in the management science* (pp. 121-153). Amsterdam: Elsevier.

Burwell, T., Jarvis, J., & McKnew, M. (1993). Modeling co-located servers and dispatch ties in hypercube model.

Computers & Operations Research, 20(2), 113-119. http://dx.doi.org/10.1016/0305-0548(93)90067-S.

Chelst, K., & Barlach, Z. (1981). Multiple unit dispatches in emergency services: models to estimate system performance. *Management Science*, 27(12), 1390-1409. http://dx.doi.org/10.1287/mnsc.27.12.1390.

Chiyoshi, F., Galvão, R., & 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., Iannoni, A., & Morabito, R. (2011). A tutorial on hypercube queueing models and some practical applications in emergency service systems. *Pesquisa Operacional*, 31(2), 271-299. http://dx.doi.org/10.1590/S0101-74382011000200005.

Departamento Nacional de Infraestrutura de Transportes – DNIT. (2015). *Relatório de gestão customizado: Exercício 2014*. Brasília: DNIT.

Ghussn, L., & Souza, R. (2016). Análise de desempenho do SAMU-Bauru/SP em períodos de pico de demanda. *GEPROS*, 11(3), 75-103.

Iannoni, A. P., & Morabito, R. (2006). Modelo de fila hipercubo com múltiplo despacho e backup parcial para análise de sistemas de atendimento médico emergenciais em rodovias. *Pesquisa Operacional*, 26(3), 493-519. http://dx.doi.org/10.1590/S0101-74382006000300004.

Iannoni, A. P., Chiyoshi, F., & Morabito, R. (2015). A spatially distributed queuing model considering dispatching policies with server reservation. *Transportation Research Part E, Logistics and Transportation Review*, 75, 49-66. http://dx.doi.org/10.1016/j.tre.2014.12.012.

Iannoni, A., & Morabito, R. (2008). A multiple dispatch and partial backup hypercube queueing model to analyze emergency medical systems on highways. *Transportation Research Part E, Logistics and Transportation Review*, 43(6), 755-771. http://dx.doi.org/10.1016/j.tre.2006.05.005.

Iannoni, A., 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.

JCNet. (2010, 22 de abril). *Em Bauru, Samu regional vai integrar 16 cidades*. Novo Hamburgo: Revista Emergência. Recuperado em 14 de maiode 2016, de http://www.revistaemergencia.com.br/site/content/noticias/noticia_detalhe.php?id=AJy4Jy

Larson, R. (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., & Odoni, A. (2007). *Urban operations research*. Prentice-Hall. Recuperado em 14 de maiode 2016, de http://web.mit.edu/urban_or_book/www/book/

Little, J. D. (2011). OR FORUM: little's law as viewed on its 50th anniversary. *Operations Research*, 59(3), 536-549. http://dx.doi.org/10.1287/opre.1110.0940.

Lopes, S. L., & Fernandes, R. J. (1999). Uma breve revisão do atendimento médico pré-hospitalar. *Medicina,*, 32, 381-387.

Organização Mundial da Saúde – OMS. (2016). *Global health observatory*. Recuperado em 5 de outubro de 2016, de http://www.who.int/gho/en/

Rodrigues, L. (2014). *Análise dos serviços emergenciais de manutenção agrícola e barracharia na agroindústria canavieira utilizando teoria das filas* (Tese de doutorado). Universidade Federal de São Carlos, São Carlos.

Sacks, S., & Grief, S. (1994). *Orlando Police Department uses OR/MS methodology, new software to design patrol disctricts* (pp. 30-32). Baltimore: OR/MS Today.

Schmid, V. (2012). Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. *European Journal of Operational Research*, 219(3), 611-621. PMid:25540476. http://dx.doi.org/10.1016/j.ejor.2011.10.043.

Simpson, N., & Hancock, P. (2009). Fifty years of operational research and emergency response. *The Journal of the Operational Research Society*, 60(S1), S126-S139. http://dx.doi.org/10.1057/jors.2009.3.

Souza, R. M. (2010). *Análise da configuração de SAMU utilizando modelo hipercubo com prioridade na fila e múltiplas alternativas de localização de ambulâncias* (Tese de doutorado). Universidade Federal de São Carlos, São Carlos.

Souza, R. M., Morabito, R., Chiyoshi, F. Y., & Iannoni, A. P. (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.

Souza, R. M., Morabito, R., Chiyoshi, F. Y., & Iannoni, A. P. (2014). Extensão do modelo hipercubo para análise de sistemas de atendimento médico emergencial com prioridade na fila. *Produção*, 24(1), 1-12. http://dx.doi.org/10.1590/S0103-65132013005000028.

Souza, R., Morabito, R., Chiyoshi, F., & Iannoni, A. (2015). Incorporating priorities for waiting customers in the hypercube queuing model with application to an emergency medical service system in Brazil. *European Journal of Operational Research*, 242(1), 274-285. http://dx.doi.org/10.1016/j.ejor.2014.09.056.

Swersey, A. (1994). The deployment of police, fire, and emergency medical units. In S. Pollock, M. Rothkopf & A. Barnett, *Handbooks in OR and MS* (pp. 151-200). Amsterdam: Elsevier.

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 médio de urgência. *Pesquisa Operacional*, 24(1), 39-71. http://dx.doi.org/10.1590/S0101-74382004000100004.

Takeda, R., Widmer, J., & Morabito, R. (2007). Analysis of ambulance decentralization in an urban emergency medical service using the hypercube queueing model. *Computers & Operations Research*, 34(3), 727-741. http://dx.doi.org/10.1016/j.cor.2005.03.022.