Recebido: 01/09/2015
Aprovado: 20/02/2016

Análise de desempenho do SAMU/Bauru-SP em períodos de pico de demanda

Performance analysis of the SAMU/Bauru-SP in peak demand periods

Larissa de Souza Ghussn[1] - Fac. de Engenharia de Bauru - Dep. de Engenharia de Produção
Regiane Máximo de Souza[2] - Fac. de Engenharia de Bauru - Dep. de Engenharia de Produção

RESUMO A qualidade de vida da população está diretamente ligada ao acesso à saúde. Uma vez ocorrido um chamado de urgência ou emergência aos Serviços de Atendimentos Emergenciais (SAE's), a resposta rápida a tal demanda é fundamental para minimizar possíveis sequelas decorrentes no quadro dos pacientes. Um SAE importante para a sociedade é o Serviço de Atendimento Móvel de Urgência (SAMU/192), que é um programa do governo federal e tem como finalidade prestar socorro à população em casos de urgência e emergência. Nesse contexto, o tempo médio de resposta ao usuário é de fundamental importância, pois a demora no atendimento pode significar a vida ou a morte de uma pessoa. Devido às restrições orçamentárias, os SAE's não podem ter um grande número de pessoas e equipamentos. Dessa forma, existe um compromisso evidente entre investimentos, custos operacionais e o nível de serviço oferecido aos usuários. Uma forma objetiva de avaliar o sistema principalmente em períodos de alta demanda é importante para os gestores, a fim de dar subsídios para tomada de decisão. Nesse sentido, os objetivos desse trabalho são: (i) descrever os chamados e os atendimentos do SAMU-Bauru/SP, (ii) aplicar o modelo hipercubo e (iii) obter suas principais medidas de desempenho. Para isso, realizou-se um estudo de caso no Serviço de Atendimento Móvel de Urgência (SAMU) no município de Bauru. A partir dessa análise, foi possível obter as principais medidas de desempenho do SAMU-Bauru/SP.

Palavras-chave SAMU. Análise de dados. Medidas de desempenho. Modelo hipercubo.

ABSTRACT *The population's quality of life is linked to access to health care services. In the event of urgent or emergency calls to Emergency Services (SAE's), rapid response to this demand is essential to minimize the possible consequences of the patient's condition. The Mobile Emergency Service (SAMU/192) is an important SAE to society. It is a federal program and provides the population with emergency care. In this context, the average response time is of fundamental importance because a delay in treatment can mean life or death. Due to budget constraints, the SAE's cannot have a large number of staff or equipment. Thus, there is a clear trade-off between investment, operating costs and the service level. An objective form to evaluate the system, especially in periods of high demand, is important for managers in order to make allowances for decision-making. In this sense, the objectives of this study are: (i) to describe the calls to the SAMU-Bauru/SP, (ii) to apply the hypercube model and (iii) to obtain key performance measures. To this end, we conducted a case study of the Mobile Emergency Service (SAMU) in Bauru. From this analysis, we it was possible to obtain the key performance measures of the SAMU- Bauru/SP.*

Keywords *SAMU. Data analysis. Performance measures. Hypercube model.*

1. Av. Engenheiro Luís Edmundo Carrijo Coube, 14-01, Vargem Limpa, CEP 17033-360, Bauru, SP, larrissaghussn@gmail.com
2. regiane@feb.unesp.br

GHUSSN, L. S.; SOUZA, R. M. Análise de desempenho do SAMU/Bauru-SP em períodos de pico de demanda. **GEPROS. Gestão da Produção, Operações e Sistemas**, Bauru, Ano 11, nº 3, jul-set/2016, p. 75-103.

DOI: 10.15675/gepros.v11i3.1460

1. INTRODUÇÃO

A qualidade de vida da população está diretamente ligada ao acesso à saúde. A organização dos Sistemas de Atendimentos Emergenciais (SAE's) ganham maior relevância e causam forte impacto ao setor saúde no Brasil, uma vez que o país ocupou o 5º lugar somente em mortes no trânsito no mundo. A resposta rápida a tal demanda é fundamental para minimizar possíveis sequelas decorrentes no quadro dos pacientes (SOUZA, 2010).

Os Sistemas de Serviços de Atendimento Emergenciais (SAE´s) são projetados e operados com o objetivo de atender a população com o menor tempo de resposta possível, considerando as limitações dos seus recursos. Para reduzir o tempo de resposta, a maioria dos sistemas de emergência adaptam mudanças nas condições de utilização de "gestão de status do sistema" – um conjunto de estratégias que incluem o reposicionamento dinâmico, no qual se modifica a localização de ambulâncias a fim de obter uma maior cobertura na área de atendimento (ALANIS; INGOLFSSON; KOLFAL, 2012).

Nesse contexto, o tempo médio de resposta ao usuário é de fundamental importância, pois a demora no atendimento pode significar a vida ou a morte de uma pessoa. Devido às restrições orçamentárias, os SAE's não podem ter um grande número de pessoas e equipamentos, com mais ambulâncias e tripulações. Portanto, existe um compromisso (*trade-off*) evidente entre investimentos, custos operacionais e o nível de serviço oferecido aos usuários (SOUZA, 2010).

No Brasil, o governo federal adotou um tipo de serviço de atendimento emergencial conhecido como Serviço de Atendimento Móvel de Urgência (SAMU/192), que teve início por meio de um acordo bilateral, assinado entre o Brasil e a França em que as viaturas de suporte avançado possuem, obrigatoriamente, a presença de um médico e tem como finalidade prestar socorro à população em casos de emergência (LOPES; FERNANDES, 1999). Esse serviço funciona 24 horas por dia com equipes de profissionais de saúde como médicos, enfermeiros, auxiliares de enfermagem e socorristas que atendem às urgências de natureza traumática, clínica, pediátrica, cirúrgica, gineco-obstétrica e de saúde mental da população. O SAMU realiza o atendimento de urgência e emergência em locais como: residências, locais de trabalho e vias públicas. O socorro é feito após chamada gratuita feita para o telefone 192 (SAMU-192). A demanda de usuários do SAMU em uma região urbana é, usualmente, separada por sub-regiões e classes de chamados de emergência. Essa demanda pode mudar significativamente ao longo do dia, geográfica e temporalmente, devido à sua natureza aleatória, mas também devido aos diferentes padrões de comportamento da população ao longo do dia.

O modelo hipercubo, proposto originalmente por Larson (1974) tem se mostrado eficiente e preciso para analisar SAE´s como, por exemplo, nos Estados Unidos, em Chelst e Barlach (1981), Brandeau e Larson (1986), Burwell et al. (1993), Sacks e Grieff (1994), Swersey (1994) e Larson e Odoni (2007). No Brasil, alguns exemplos aparecem em Gonçalves et al. (1995), Mendonça e Morabito (2000), Oliveira (2003), Chiyoshi et al. (2000), Costa et al. (2004), Figueiredo et al. (2005), Takeda et al. (2004, 2007) e Iannoni (2005). A aplicação original do modelo hipercubo foi desenvolvida para o problema de patrulhamento policial, mas depois, o modelo passou a ser aplicado em vários sistemas de emergência como, empresas de segurança, bombeiros, ambulâncias, reparos em redes de energia elétrica, entre outros (LARSON; ODONI, 2007).

Algumas medidas de desempenho mais relevantes para um sistema de atendimento de urgência podem ser divididas em: medidas externas, do ponto de vista do usuário, como o tempo médio de resposta a um chamado, o tempo médio de viagem para cada área da cidade e a frequência de chamadas atendidas em um tempo inferior a um limite determinado; e medidas internas, do ponto de vista do gerente do sistema, como a carga de trabalho das ambulâncias, as frequências de despacho das ambulâncias para os átomos, a fração de atendimentos realizados fora da área de cobertura de cada ambulância e o tempo médio de viagem para cada ambulância. O modelo hipercubo tem por objetivo avaliar a configuração e estimar as medidas de desempenho para SAE's, como, por exemplo, os SAMU's, possibilitando um planejamento adequado e melhores níveis de serviço oferecido (GALVÃO; MORABITO, 2008)

Dada a importância dos SAMU´s nas cidades brasileiras, os objetivos desse trabalho são: (i) descrever os chamados e os atendimentos do SAMU-Bauru/SP, (ii) aplicar o modelo hipercubo e (iii) obter suas principais medidas de desempenho. Para isso, realizou-se um estudo de caso no Serviço de Atendimento Móvel de Urgência (SAMU) no município de Bauru. Para isso, realizou-se um estudo de caso no Serviço de Atendimento Móvel de Urgência (SAMU) no município de Bauru, onde se localiza sua base, através da análise estatística dos chamados e atendimentos.

Na próxima Seção, será apresentada uma descrição do SAMU – Bauru/SP. A Seção 3 apresentará a descrição dos atendimentos do sistema. A Seção 4 mostrará a validação das hipóteses para aplicação do modelo hipercubo. A Seção 5 apresentará a aplicação do modelo hipercubo no SAMU/Bauru e a Seção 6 irá apresentar as conclusões desse estudo.

2. MÉTODO

2.1. O cenário do estudo: O SAMU – Bauru/SP

O SAMU – Bauru/SP, onde o estudo se realizou, integra 16 cidades da região, numa parceria com suas prefeituras, tendo suas bases instaladas em sete delas. A regionalização abrange as cidades de Bauru, cidade sede, que se subdivide em mais regiões, tendo diariamente sete ambulâncias básicas e três avançadas em atendimento; Pederneiras, Lençóis Paulista, Agudos, Arealva, Pirajuí, Duartina, as quais possuem bases com uma ambulância básica; Macatuba, Borebi, Reginópolis, Presidente Alves, Cabrália Paulista, Lucianópolis e Avaí (Figura 1).

Figura 1 – Mapa das cidades atendidas pelo SAMU – Bauru/SP.

Fonte: Elaboração das autoras.

Há dois tipos de ambulâncias no SAMU – Bauru/SP. Os veículos de suporte básico (VSB's) caracterizam-se por ter um motorista e um auxiliar de enfermagem, responsáveis pelo atendimento a pacientes em casos de baixo risco. Já os veículos de suporte avançado (VSA's) possuem um motorista, um auxiliar de enfermagem, um enfermeiro e um médico, atendendo, portanto, aos chamados mais graves. Essas se localizam apenas no município de Bauru, sendo duas destinadas ao atendimento local e uma voltada para o atendimento regional.

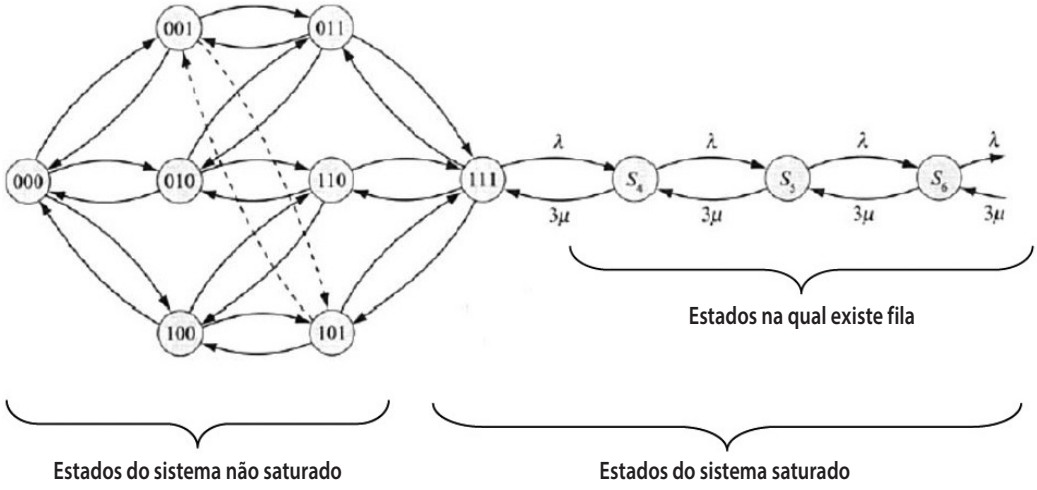
2.2. O modelo hipercubo

O modelo hipercubo é um modelo analítico em que a ideia é expandir o espaço de estados de um sistema de fila M/M/m (notação de Kendal, em que o primeiro M indica que a distribuição do intervalo entre chegadas é uma exponencial, o segundo M indica que a distribuição dos tempos de serviço também é uma exponencial e o sistema possui m servidores) a fim de representar cada servidor individualmente, podendo considerar políticas de despacho mais complicadas no sistema. A solução do modelo é dada partindo-se da construção de um conjunto de equações de equilíbrio (*steady state*) para o sistema. Os resultados baseiam-se nos valores das probabilidades de estado de equilíbrio do sistema, possibilitando o cálculo de diversas medidas de desempenho, tais como: cargas de trabalho dos servidores, tempo médio de resposta do sistema ou de cada servidor, frequência de atendimento de cada servidor em cada região, entre outras.

O modelo baseia-se na divisão da região atendida pelo sistema em átomos geográficos (regiões de demanda). Cada átomo é considerado uma fonte de chamados pontual e independente das demais e o atendimento a cada átomo é realizado por servidores que estão distribuídos na região. A localização dos servidores deve ser conhecida ou, caso contrário, estimada (LARSON; ODONI, 2007). Se um servidor estiver ocupado, outros servidores poderão atender ao chamado, mesmo que estejam fora de sua região preferencial do chamado, prevalecendo a cooperação entre os servidores. Para considerar prioridade no despacho de ambulâncias quando o sistema não se encontra saturado, é realizada a estratégia de "*layering*" (LARSON; ODONI, 2007) que consiste na divisão dos átomos geográficos em subátomos.

A disponibilidade dos servidores é representada por meio do espaço de estados dos servidores. Um estado particular do sistema sem fila é dado pela lista dos servidores que estão livres ou ocupados. Considere um sistema com apenas m = 3 servidores e sejam {000}, {001}, {010},..., {111} os $2^3 = 8$ possíveis estados do sistema, em que os 0's e 1's indicam se cada um dos três servidores está livre ou ocupado, respectivamente. Por exemplo, o estado {011} representa o estado em que o servidor 1 está livre e os servidores 2 e 3 estão ocupados (note que {011} descreve o estado dos servidores da esquerda para a direita). Assim, o espaço de estados desse sistema com três servidores pode ser representado por um cubo; no caso de haverem mais de três servidores, temos um hipercubo. A Figura 2 ilustra o espaço de estados desse sistema com três servidores. O modelo hipercubo trata tanto sistemas em que não é permitida a formação de fila, como aqueles em que quando todos os servidores estão ocupados, os chamados que chegam esperam em uma fila por meio da qual os usuários são atendidos à medida que os servidores tornam-se desocupados segundo a disciplina FCFS (*First Come First Served*). Os demais estados S_4, S_5, S_6, ..., da Figura 1 representam os estados com 1, 2, 3, ..., usuários na fila de espera do sistema, respectivamente.

Figura 2 – Estados do sistema com três servidores.



Fonte: Adaptado de Larson & Odoni, 2007.

No caso dos m servidores serem homogêneos, com mesma taxa de atendimento, o modelo hipercubo, que considera os m servidores distinguíveis, possui uma relação direta com o modelo clássico de fila M/M/m, que considera os m servidores indistinguíveis. Segundo Larson e Odoni (2007), existem nove hipóteses principais que devem ser verificadas para a aplicação do modelo hipercubo:

i. A região deve ser dividida em N_A átomos.
ii. As solicitações por serviço em cada átomo j (j = 1, 2, ... , N_A) chegam independentemente e de acordo com uma distribuição de Poisson.
iii. Os tempos de viagem de um átomo i para um átomo j (i, j = 1, ... , N_A) devem ser conhecidos ou estimados.
iv. O sistema opera com m servidores espacialmente distribuídos, homogêneos ou não, que podem se deslocar para atender qualquer um dos átomos.
v. Quando disponíveis, a localização dos servidores deve ser conhecida, ao menos probabilisticamente.
vi. Apenas um servidor é despachado para atender cada chamado. Quando chamados estiverem esperando em fila, a escolha do chamado a ser atendido utiliza a disciplina FCFS.
vii. Há uma lista de preferências de despacho de servidores para cada átomo.
viii. O tempo total de atendimento de um chamado é exponencialmente distribuído e composto pela somatória dos seguintes tempos: tempo de preparo do servidor (*setup time*), tempo de viagem do servidor até o local da ocorrência, tempo de execução do serviço junto ao usuário (tempo em cena) e o tempo de viagem de retorno à base.
ix. Variações no tempo total de atendimento devido às variações no tempo de viagem são consideradas de segunda ordem, quando comparadas às variações dos tempos em cena e/ou tempo de preparação da equipe.

Como enfatizado antes, algumas dessas hipóteses podem ser modificadas ou desconsideradas. A seguir, é apresentado o modelo hipercubo por meio de um exemplo simples, resolvido em Chiyoshi et al. (2001). Considere um sistema de emergência operando em uma região representada por três átomos, utilizando política de despacho de preferência fixa, mostrada no Quadro 1.

Quadro 1 – Matriz de Preferências de despacho.

Átomo	Matriz de Despachos		
	Preferências		
	1º	2º	3º
1	1	2	3
2	2	3	1
3	3	1	2

Fonte: Elaboração das autoras.

A solução do modelo é dada pela construção das equações de equilíbrio do sistema, que são definidas supondo-se que o sistema atinja o equilíbrio. Para cada estado do sistema, o fluxo que entra neste estado deve ser igual ao fluxo que sai dele. Em um sistema não saturado, com capacidade de fila infinita, as probabilidades de estado do modelo hipercubo são calculadas a partir das equações de balanço, construídas a partir dos oito possíveis estados, descritos anteriormente nesta mesma seção.

Quando o sistema está no estado {000} (sistema vazio, Figura 2), ele passa para o estado {100} quando ocorre um chamado com origem no átomo 1, com taxa de ocorrência λ_1. O mesmo acontece com o estado {010}, com taxa λ_2, e para o estado {001}, com taxa λ_3. Dessa forma, a taxa total de transição do estado {000} para outros estados é $\lambda = \lambda_1 + \lambda_2 + \lambda_3$.

No sentido contrário, Figura 2, o estado {000} pode ser alcançado a partir do estado {100} quando o servidor 1 termina o atendimento, com taxa μ_1; da mesma forma, a partir do estado {010}, com taxa μ_2; e a partir de {001}, com taxa μ_3. Podemos obter a equação de equilíbrio para o estado {000} a partir da definição de que "a taxa com que o sistema entra no estado n deve ser igual a taxa com que o sistema sai do estado n" (SOUZA, 2010), da seguinte forma:

$$\lambda P_{\{000\}} = \mu_1 P_{\{100\}} + \mu_2 P_{\{010\}} + \mu_3 P_{\{001\}} \quad (1)$$

Com esse procedimento, podemos obter as equações para os estados seguintes, obtendo o conjunto de Equações 2.

$$
\begin{aligned}
&\{000\} \rightarrow \lambda P_{\{000\}} = \mu_1 P_{\{100\}} + \mu_2 P_{\{010\}} + \mu_3 P_{\{001\}} \\
&\{001\} \rightarrow (\lambda + \mu_1) P_{\{100\}} = \lambda_1 P_{\{000\}} + \mu_2 P_{\{110\}} + \mu_3 P_{\{101\}} \\
&\{010\} \rightarrow (\lambda + \mu_2) P_{\{010\}} = \lambda_2 P_{\{000\}} + \mu_1 P_{\{110\}} + \mu_3 P_{\{011\}} \\
&\{100\} \rightarrow (\lambda + \mu_3) P_{\{001\}} = \lambda_3 P_{\{000\}} + \mu_1 P_{\{101\}} + \mu_2 P_{\{011\}} \\
&\{011\} \rightarrow (\lambda + \mu_1 + \mu_2) P_{\{110\}} = (\lambda_1 + \lambda_2) P_{\{100\}} + \lambda_1 P_{\{010\}} + \mu_3 P_{\{111\}} \\
&\{101\} \rightarrow (\lambda + \mu_1 + \mu_3) P_{\{101\}} = \lambda_3 P_{\{100\}} + (\lambda_1 + \lambda_3) P_{\{001\}} + \mu_2 P_{\{111\}} \\
&\{110\} \rightarrow (\lambda + \mu_2 + \mu_3) P_{\{011\}} = (\lambda_2 + \lambda_3) P_{\{010\}} + \lambda_2 P_{\{001\}} + \mu_1 P_{\{111\}} \\
&\{111\} \rightarrow (\lambda + \mu) P_{\{111\}} = \lambda P_{\{110\}} + \lambda P_{\{101\}} + \lambda P_{\{011\}} + \mu P_{\{S_4\}}
\end{aligned}
\tag{2}
$$

Em que:

- λ_i é a taxa de chegada de chamadas no átomo i;
- μ_j é a taxa de atendimento do servidor j;
- $\lambda = \lambda_1 + \lambda_2 + \lambda_3$ é a taxa total de chegada no sistema;
- $\rho = \frac{\lambda}{\mu}$ é a carga média de trabalho no sistema.

Pela condição de equilíbrio do sistema, a transição entre os estados {111} e $\{S_4\}$ devem ser iguais, de forma que $\lambda P111 = \mu P_{S_4}$. Caso essa condição não aconteça, o sistema está na fase transiente e a cauda estaria em crescimento. Assim, a oitava equação do Sistema 2 pode ser escrita na forma:

$$\{111\} \rightarrow (\lambda + \mu) P_{\{111\}} = \lambda P_{\{110\}} + \lambda P_{\{101\}} + \lambda P_{\{011\}} + \mu P_{\{S_4\}}$$

Chiyoshi et al. (2000) mencionam que o sistema 2 escrito na forma matricial $Ax = 0$ é um sistema linear homogêneo indeterminado. Isso ocorre porque as equações apenas impõem condições de equilíbrio para cada estado do sistema {000},{001},{010},{100},{011},{101},{110},{111}, mas nada especifica sobre como a massa total de probabilidade se distribui entre estes estados e os estados da cauda. Uma maneira de tornar o sistema determinado é a substituição de uma das equações do sistema por uma equação de normalização, considerando que $\sum_{n=0}^{N} P_B = 1$, sendo que N é o número de estados possíveis para o sistema. A Equação de normalização é dada por:

$$P_{\{000\}} + P_{\{001\}} + P_{\{010\}} + P_{\{100\}} + \ldots + P_{\{111\}} + P_{\{S_4\}} + P_{\{S_5\}} + \ldots = 1 \tag{3}$$

A solução do Sistema de Equações 2 resulta nas probabilidades dos estados do sistema. Diversas medidas de desempenho podem ser calculadas a partir dessas probabilidades de o sistema estar em cada estado, obtidas pela solução dessas equações de equilíbrio do sistema. Essas medidas auxiliam na configuração e análise do sistema sob a hipótese de que o sistema esteja em equilíbrio (LARSON; ODONI, 2007).

2.3. Medidas de desempenho

As medidas de desempenho podem ser calculadas de diferentes maneiras, conforme a existência ou não de fila no sistema e se a restrição no atendimento é física ou não. Neste trabalho, apresentar-se-á apenas as medidas para sistemas onde ocorrem filas com prioridade e sem restrições físicas para o atendimento.

Para sistemas de *backup* parcial sem fila, o cálculo das medidas de desempenho pode ser encontrado em detalhes no trabalho de Iannoni (2005). Já para sistemas com filas e restrição física no atendimento tem-se os cálculos presentes em estudos de Rodrigues (2014).

Primeiramente, a carga de trabalho, ρ_n (Equação 4), para cada servidor (n) pode ser obtida pela soma das probabilidades dos estados em que o dado servidor está ocupado (LARSON; ODONI, 2007):

$$\rho_i = \sum_{\{B:b_i=1\}} P_B + P_Q \text{ , em que:} \tag{4}$$

- ρ_i é a carga de trabalho (*workload*) do servidor i (i = 1, 2, ..., m);

- $\sum_{\{B:b_i=1\}} P_B$ é a soma das probabilidades dos estados (de {000} a {111}), em que o servidor i está ocupado ($b_i = 1$);

- P_Q é a probabilidade de fila ($P_Q = P_{\{s_4\}} + P_{\{s_5\}} + ...$).

Seguindo com as frequências de despacho, f_{ijk}, dos servidores i aos subátomos jk, são calculadas somando-se dois termos. Sendo o primeiro termo f_{ijk}^{nq}, que é a fração de todos os despachos do servidor i para o subátomo jk que não incorre em fila. E o segundo termo é f_{ijk}^{nq}, que mostra a fração de todos os despachos do servidor i ao subátomo jk que incorre em fila (Equação 5).

$$f_{ijk} = f_{ijk}^{[nq]} + f_{ijk}^{[q]} = \underbrace{\frac{\lambda_{jk}}{\lambda} \sum_{B \in E_{ijk}} P_B}_{f_{ijk}^{[nq]}} + \underbrace{\frac{\lambda_{jk}}{\lambda} P_{Qk'} \frac{\mu_i}{\mu}}_{f_{ijk}^{[q]}} \tag{5}$$

Em que f_{ijk}^{nq}, é a frequência de despacho dos chamados em que não ocorrem fila no sistema e f_{ijk}^{q} é a frequência de despacho dos chamados em que ocorrem fila no sistema. E_{ijk} é o conjunto dos estados nos quais o servidor i é o primeiro servidor disponível na lista de despacho do subátomo jk. P_{Qk} é a probabilidade de saturação do sistema ($P_{Qk'} = P_{Qk} + P_{\{111\}}$). P_{Qk} é a probabilidade de saturação do sistema menos a probabilidade de todos os servidores estarem ocupados, isto é, a probabilidade de fila do sistema. Além disso, em sistemas que possuem perda, a taxa de entrada de usuários é dada por pela Equação 6.

$$\bar{\lambda} = \lambda(1 - P_K) \tag{6}$$

Em que K é a capacidade do sistema.

Para calcular o tempo médio de viagem para o sistema é necessário conhecer a localização dos servidores, o tempo médio necessário para um servidor m, quando disponível, viajar até o subátomo jk, e o tempo médio de espera de um chamado que está em fila.

A representação da localização dos servidores é feita a partir de uma matriz $L = [l_{i,jk}]$, em que os elementos representam a probabilidade de um servidor i estar localizado em um subátomo jk, quando disponível. Sendo L uma matriz estocástica, ou seja $\sum_{j=1}^{N_A}\sum_{k\in D} l_{i,jk} = 1$, se o servidor i está localizado no subátomo jk, então $[l_{i,jk}] = 1$, e $[l_{i,jk}] = 0$ se o servidor i não está localizado no subátomo jk. Assim, o tempo médio de viagem para um servidor se deslocar até um determinado subátomo é dado pela Equação 7.

$$t_{i,pl} = \sum_{j=1}^{N_A}\sum_{k\in D} l_{i,jk} \cdot \tau_{jk,pl} \tag{7}$$

O tempo médio de viagem para cada subátomo jk ($\bar{T}_{jk}$) é outra medida que reflete o nível de serviço oferecido pelo sistema (SOUZA, 2010). A Equação 8 dá o tempo médio de viagem ao subátomo jk utilizando a disciplina FCFS.

$$\bar{T}_{jk} = \frac{\sum_{i=1}^{m} f_{i,jk}^{[nq]} t_{i,jk}^{[nq]}}{\sum_{i=1}^{m} f_{i,jk}^{[nq]}}\left(1 - P_{Q'}\right) + \sum_{p=1}^{N_A}\sum_{l\in D}\frac{\lambda_{pl}}{\lambda}\tau_{pl,jk}P_{Q'} \tag{8}$$

Em um sistema em que há usuários separados em classes, é interessante calcular o tempo médio de viagem de cada servidor para cada classe k, esta medida pode ser calculada a partir da expressão 9.

$$\overline{TU}_{ik} = \frac{\sum_{j=1}^{N_A} f_{i,jk}^{[nq]} t_{i,jk}^{[nq]} + \sum_{j=1}^{N_A} f_{i,jk}^{[q]} t_{i,jk}^{[q]}}{\sum_{j=1}^{N_A} f_{i,jk}^{[nq]} + \sum_{j=1}^{N_A} f_{i,jk}^{[q]}} \quad (9)$$

O tempo de espera para chamados em fila (Equação 10) pode ser calculado utilizando a fórmula de Little (LITTLE, 1961). Se definirmos o número de usuários na fila de uma determinada classe r como n_r, a probabilidade $P(n_r = j)$ é dada pela soma das probabilidades associadas com os estados na qual os usuários da classe r está em j (SOUZA et al. 2015).

$$P(n_r = j) = \sum_{\forall S \; s.t. \; n(r,S)=j} P\{S\} \quad (10)$$

Onde S é um estado da fila e n(r,S) é o número de usuários da classe r em S. Desta distribuição, o número médio de usuários da classe k pode ser determinado pela Equação 11:

$$L_{qr} = \sum_j jP(n_r = j) \quad (11)$$

Do qual o tempo médio de espera é obtido pela Equação 12:

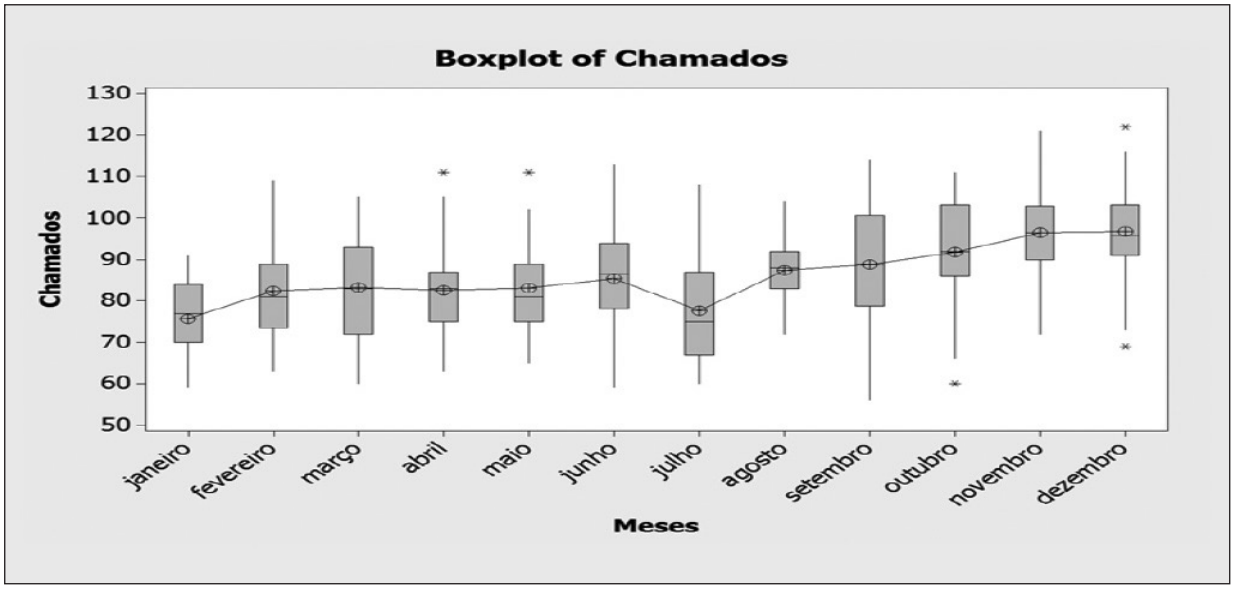
$$W_{qr} = L_{qr}/\lambda_r \quad (12)$$

3. DESCRIÇÃO DOS ATENDIMENTOS DO SISTEMA

A descrição dos atendimentos foi realizada como em Takeda (2000), Takeda et al. (2007), Souza (2010) e Souza et al. (2015). Todos os chamados recebidos na base do SAMU – Bauru/SP passam por um médico regulador, que indica as providências a serem tomadas imediatamente com o paciente até que a ambulância chegue. Os chamados são divididos por cores (azul, verde, amarelo e vermelho, do mais leve para o mais grave, respectivamente), indicando o tipo de risco e o tempo de espera do paciente, possibilitando que um chamado mais grave seja atendido com prioridade em relação a um chamado de baixo risco.

Os dados foram coletados por meio dos relatórios de síntese de atendimento, constituídos de dados secundários disponíveis no SAMU – Bauru/SP. A coleta de dados foi realizada em duas fases. Primeiramente, fez-se um levantamento do número de atendimentos, entre janeiro de 2012 a fevereiro de 2013, fornecido pelos gestores do SAMU – Bauru/SP a fim de verificar se havia diferenças estatísticas significativas no número de atendimentos em cada mês do ano analisado. Também foi feito o Boxplot dos chamados (Figura 3). Além disso, verificou-se a existência de diferenças com relação ao número de chamados das semanas. Para verificar se o número médio de atendimentos é estatisticamente igual em todos os meses, aplicou-se a Análise de Variância (ANOVA), com nível de significância $\alpha = 0,05$. O banco de dados foi organizado pelo software Microsoft Excel® e a análise estatística dos dados, feita pelo software Minitab®.

Figura 3 – Boxplot da quantidade de chamados por mês de 2012.

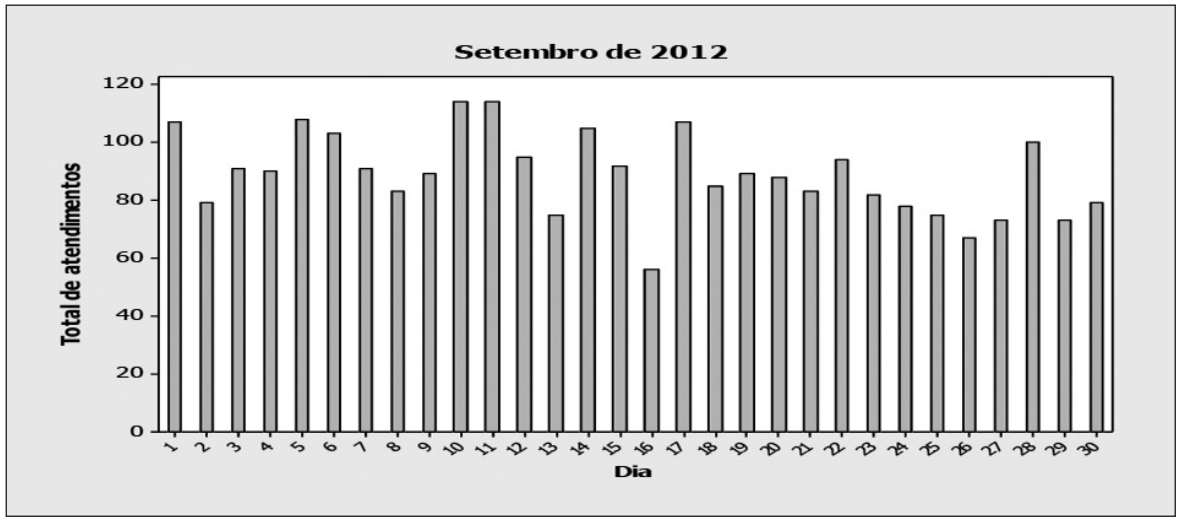


Fonte: Elaboração das autoras.

Verificou-se que há diferença entre os meses e, para identificar os meses diferentes, foi aplicado o teste de Tukey. Verificou-se que os meses Ago./12, Set./12, Out./12, Nov./12, dez/12. Jan/13 e Fev./13 foram agrupados como sendo estatisticamente iguais em relação à quantidade de chamados e com maior número de chamados atendidos nos meses analisados. Desses meses observados, foi sorteado o mês de setembro de 2012 para fazer a segunda fase de coleta de dados. A Figura 4 mostra a frequência dos chamados desse mês.

No mês de Setembro de 2012, não foram verificadas diferenças significativas em relação à quantidade de chamados nas quatro semanas do mês, por meio da Análise de Variância (ANOVA) com nível de significância $\alpha = 0,05$.

Figura 4 – Total de atendimentos por dia do mês de Setembro de 2012.

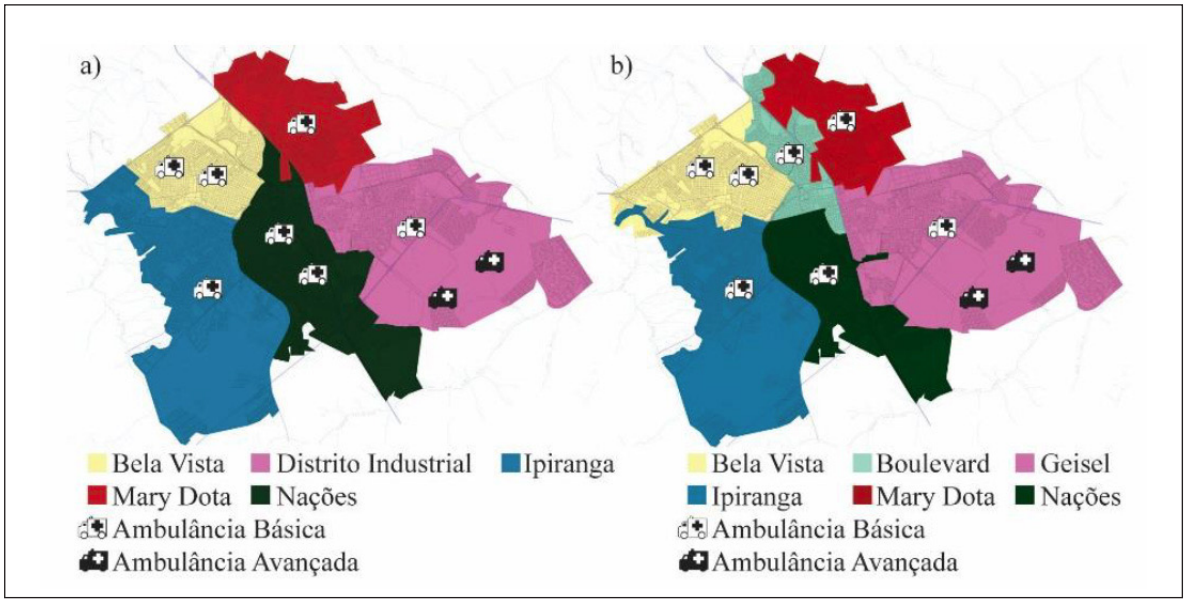


Fonte: Elaboração das autoras.

Sortearam-se 10 dias do mês de setembro de 2012 para proceder com a segunda fase da coleta de dados. Para cada um dos dez dias escolhidos, durante todo o dia, realizou-se um levantamento minucioso dos chamados, anotando: o horário do chamado, a região de origem, o tipo de urgência, a ambulância que o atendeu, o tempo de envio de equipe e saída da base e tempos de viagens dos servidores.

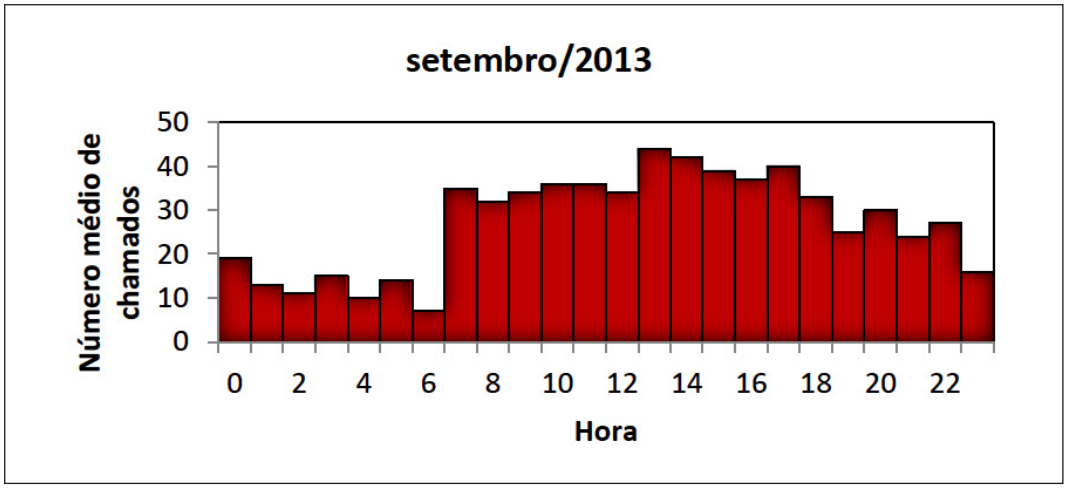
No decorrer da coleta de dados, foram constatadas mudanças significativas no atendimento do SAMU – Bauru/SP. Houve alterações nas bases municipais (Figura 5) e também ocorreu um aumento no número de chamados atendidos pelas unidades avançadas, as quais passaram a atender aos chamados graves com o auxílio de motos, agilizando o contato com o paciente. Devido a esses fatores, também foi realizada a coleta de dados em dez dias do mês de setembro de 2013. A partir dos dados obtidos na segunda fase da coleta de dados, procedeu-se a identificação do (s) período(s) de pico durante o dia, escolhido(s) de forma que apresente(m) maior taxa média de chegada e menor desvio-padrão a partir dos dados coletados em setembro de 2013, que é a configuração atual do SAMU – Bauru/SP. A Figura 6 mostra o número médio de chamados por hora dos 10 dias coletados de setembro de 2013.

Figura 5 – Mapa de Bauru/SP e as representações dos átomos geográficos de 2012 (a) e 2013 (b) com as respectivas ambulâncias utilizadas pelo SAMU – Bauru/SP.



Fonte: DAE Bauru – modificado por Guilherme Ghussn.

Figura 6 – Número médio de chamados por hora dos 10 dias de coleta de dados em Setembro de 2013.



Fonte: Elaboração das autoras.

A análise dos períodos de pico foi feita com os dados obtidos a partir do controle dos chamados do SAMU – Bauru/SP. A Tabela 1 mostra a taxa de chegada (número de chamados dividido pelo período de tempo considerado em minutos) para vários períodos do dia, durante os 10 dias estudados. Foram analisados vários períodos e o período de pico em cada parte do dia foi identificado pela análise feita a partir da média e do desvio padrão da taxa de chegada. O período escolhido (linha destacada da Tabela 1) foi o que apresentou maior taxa média de chegada e menor desvio-padrão, conforme o estudo em Takeda (2000) para o SAMU-Campinas e Souza (2010) para o SAMU-RP. Alguns períodos apresentaram a média e o desvio-padrão muito próximos, assim, escolhe-se aquele de maior duração, a fim de obter uma maior quantidade de dados no período. Pode-se observar que o período escolhido foi o das 12h às 18h.

Tabela 1 – Análise do período de pico.

	DIA											
Período	1	2	3	4	5	6	7	8	9	10	média	d-p
13-18h	0,047	0,063	0,060	0,063	0,080	0,073	0,087	0,050	0,050	0,100	0,067	0,017
12-18h	0,044	0,069	0,053	0,064	0,078	0,069	0,083	0,053	0,050	0,092	0,066	0,016
13-19h	0,050	0,067	0,058	0,067	0,075	0,067	0,081	0,047	0,050	0,092	0,065	0,014
13-17h	0,042	0,071	0,067	0,063	0,096	0,075	0,079	0,050	0,042	0,092	0,068	0,019

Fonte: Elaboração das autoras.

4. VALIDAÇÃO DAS HIPÓTESES DO MODELO HIPERCUBO

Foi verificado se o sistema atende às nove hipóteses do modelo hipercubo, considerando todas as características do SAMU – Bauru/SP. A análise dos dados se encontra a seguir.

i) Área dividida em NA átomos geográficos

Há várias maneiras de se fazer a representação da área estudada em átomos geográficos, como: divisão política, bairros, setores policiais, entre outros. Neste trabalho, pretende-se utilizar a divisão por setores em 2013: Bela Vista, Boulevard, Geisel, Ipiranga, Mary Dota e Nações, utilizada pelo SAMU – Bauru/SP.

O SAMU – Bauru/SP possui classes diferenciadas de usuários do sistema, chamadas de classificação por risco e identificadas por cores (azul – a, verde – b, amarelo – c e, vermelho – d). O médico regulador é quem decide a gravidade do caso conforme a descrição do solicitante. Dessa forma, nesse trabalho, cada átomo geográfico (Nações – 1, Geisel – 2, Ipiranga – 3, Bela Vista – 4, Mary Dota – 5 e Boulevard – 6) será dividido em quatro sub-átomos (a, b, c e d), totalizando 24 sub-átomos no sistema: Nações – 1a, Geisel – 2a, Ipiranga – 3a, Bela Vista – 4a, Mary Dota – 5a, Boulevard – 6a, Nações – 1b, Geisel – 2b, Ipiranga – 3b, Bela Vista – 4b, Mary Dota – 5b, Boulevard – 6b, Nações – 1c, Geisel – 2c, Ipiranga – 3c, Bela Vista – 4c, Mary Dota – 5c, Boulevard – 6c, Nações – 1d, Geisel – 2d, Ipiranga – 3d, Bela Vista – 4d, Mary Dota – 5d, Boulevard – 6d. Assim, são devidamente representadas no modelo as quatro classes de usuários do SAMU – Bauru/SP. Conforme a Figura 4a, pode-se observar a configuração do SAMU – Bauru/SP, em 2012, com cinco átomos geográficos e, na Figura 4b, a configuração, em 2013, com seis átomos geográficos.

ii) Processo de chegada

Fez-se necessário fazer um teste de aderência nos dados para verificar estatisticamente a hipótese de processo de chegada Poisson. Os métodos utilizados foram Kolmogorov-Smirnov, Anderson Darling e Qui-Quadrado; ver Johnson et al. (1994, 1995). Para fazer a análise do processo de chegada no período de pico, foram considerados os chamados nesse período. Fez-se a análise do processo de chegada para todos os dias de observação, divididos em manhã, tarde e noite (períodos de pico), a fim de verificar se o número de chamados segue o padrão Poissoniano, uma vez que as chegadas dos chamados, em cada átomo, constituem processos de contagem com incrementos independentes. Os testes de aderência foram realizados com os chamados agregados (em todos os átomos), esses testes mostraram que, a um nível de significância de 5%, não se pode rejeitar a hipótese de que os intervalos entre chegadas sucessivas têm distribuição exponencial.

A fim de determinar a taxa média de chegada dos chamados no sistema, foram considerados os intervalos médios de chegadas sucessivas para todos os dias de observação nos períodos de pico. Pode-se verificar que, em todos os casos, os desvios-padrão são, em geral, da ordem de grandeza das médias, ou seja, os coeficientes de variação são relativamente próximos a 1. Isso é mais um indicativo de que o intervalo de tempo entre chegadas sucessivas dos chamados deve ser, de fato, exponencialmente distribuído. Considerando as devidas proporções que representam as chegadas em cada sub-átomo do sistema, admitindo que os chamados chegam independentemente e de acordo com o Processo de Poisson, encontraram-se as taxas médias $\lambda_{jk} = \lambda.p_{jk}$ ($j = 1, 2, ..., 6, k \in C$). Verifica-se, na Tabela 2, as taxas de chegadas dos chamados considerando as devidas proporções que representam as chegadas dos chamados em cada sub-átomo do sistema no período de pico. Pode-se ver na Tabela 2 as taxas de chegadas dos chamados considerando as devidas proporções que representam as chegadas dos chamados em cada subátomo do sistema no período de pico.

Tabela 2 – Taxas médias de chegada dos chamados (por hora) para cada sub-átomo.

	Átomo	Nº de chamados	p_j	$\lambda_{j,k}$ (chamados/hora)		Átomo	Nº de chamados	p_j	$\lambda_{j,k}$ (chamados/hora)
1	Nações Azul	6	0,0255	0,1004	13	Bela Vista Azul	7	0,0298	0,1172
2	Nações Verde	20	0,0851	0,3348	14	Bela Vista Verde	20	0,0851	0,3348
3	Nações Amarelo	31	0,1319	0,5189	15	Bela Vista Amarelo	26	0,1106	0,4352
4	Nações Vermelho	0	0,0000	0,0000	16	Bela Vista Vermelho	1	0,0043	0,0167
5	Geisel Azul	1	0,0043	0,0167	17	Mary Dota Azul	3	0,0128	0,0502
6	Geisel Verde	14	0,0596	0,2343	18	Mary Dota Verde	15	0,0638	0,2511
7	Geisel Amarelo	24	0,1021	0,4017	19	Mary Dota Amarelo	9	0,0383	0,1506
8	Geisel Vermelho	16	0,0681	0,2678	20	Mary Dota Vermelho	0	0,0000	0,0000
9	Ipiranga Azul	4	0,0170	0,0670	21	Boulevard Azul	2	0,0085	0,0335
10	Ipiranga Verde	14	0,0596	0,2343	22	Boulevard Verde	3	0,0128	0,0502
11	Ipiranga Amarelo	11	0,0468	0,1841	23	Boulevard Amarelo	8	0,0340	0,1339
12	Ipiranga Vermelho	0	0,0000	0,0000	24	Boulevard Vermelho	0	0,0000	0,0000
					Total		**235**	**1,0000**	**3,9333**

Fonte: Elaboração das autoras.

iii) Tempos de viagem

Fez-se necessário calcular os tempos médios de viagem de cada ambulância para cada átomo, cujos dados podem ser obtidos no próprio sistema. Mesmo que não haja dados suficientes, os tempos de viagem entre os átomos podem ser calculados. Os tempos de viagem foram obtidos a partir dos dados do próprio SAMU – Bauru/SP. Nos casos em que não foram encontradas observações do tempo de viagem entre dois átomos, calculou-se a distância entre os centroides dos átomos (a partir do software Google Earth) e, utilizando a velocidade média de 60 km/h, foi possível obter uma estimativa do tempo médio de viagem entre os átomos (indicados na tabela a seguir com "*"). A matriz dos tempos de viagem entre todos os átomos pode ser vista na Tabela 3.

Tabela 3 – Tempos médios de viagens entre subátomos.

Sub-átomos	1a	1b	1c	1d	2a	2b	2c	2d	3a	3b	3c	3d	4a	4b	4c	4d	5a	5b	5c	5d	6a	6b	6c	6d
1a	13,8	13,8	13,8	13,8	13,7	13,7	13,7	13,7	16,0	16,0	16,0	16,0	13,5	13,5	13,5	13,5	15,0*	15,0*	15,0*	15,0*	10,0	10,0	10,0	10,0
1b	13,8	13,8	13,8	13,8	13,7	13,7	13,7	13,7	16,0	16,0	16,0	16,0	13,5	13,5	13,5	13,5	15,0*	15,0*	15,0*	15,0*	10,0	10,0	10,0	10,0
1c	13,8	13,8	13,8	13,8	13,7	13,7	13,7	13,7	16,0	16,0	16,0	16,0	13,5	13,5	13,5	13,5	15,0*	15,0*	15,0*	15,0*	10,0	10,0	10,0	10,0
1d	13,8	13,8	13,8	13,8	13,7	13,7	13,7	13,7	16,0	16,0	16,0	16,0	13,5	13,5	13,5	13,5	15,0*	15,0*	15,0*	15,0*	10,0	10,0	10,0	10,0
2a	13,7	13,7	13,7	13,7	8,8	8,8	8,8	8,8	13,2	13,2	13,2	13,2	20,0	20,0	20,0	20,0	11,3	11,3	11,3	11,3	7,8	7,8	7,8	7,8
2b	13,7	13,7	13,7	13,7	8,8	8,8	8,8	8,8	13,2	13,2	13,2	13,2	20,0	20,0	20,0	20,0	11,3	11,3	11,3	11,3	7,8	7,8	7,8	7,8
2c	13,7	13,7	13,7	13,7	8,8	8,8	8,8	8,8	13,2	13,2	13,2	13,2	20,0	20,0	20,0	20,0	11,3	11,3	11,3	11,3	7,8	7,8	7,8	7,8
2d	13,7	13,7	13,7	13,7	8,8	8,8	8,8	8,8	13,2	13,2	13,2	13,2	20,0	20,0	20,0	20,0	11,3	11,3	11,3	11,3	7,8	7,8	7,8	7,8
3a	16,0	16,0	16,0	16,0	13,2	13,2	13,2	13,2	8,7	8,7	8,7	8,7	18,0*	18,0*	18,0*	18,0*	16,6	16,6	16,6	16,6	13,0*	13,0*	13,0*	13,0*
3b	16,0	16,0	16,0	16,0	13,2	13,2	13,2	13,2	8,7	8,7	8,7	8,7	18,0*	18,0*	18,0*	18,0*	16,6	16,6	16,6	16,6	13,0*	13,0*	13,0*	13,0*
3c	16,0	16,0	16,0	16,0	13,2	13,2	13,2	13,2	8,7	8,7	8,7	8,7	18,0*	18,0*	18,0*	18,0*	16,6	16,6	16,6	16,6	13,0*	13,0*	13,0*	13,0*
3d	16,0	16,0	16,0	16,0	13,2	13,2	13,2	13,2	8,7	8,7	8,7	8,7	18,0*	18,0*	18,0*	18,0*	16,6	16,6	16,6	16,6	13,0*	13,0*	13,0*	13,0*
4a	13,5	13,5	13,5	13,5	20,0	20,0	20,0	20,0	18,0*	18,0*	18,0*	18,0*	8,9	8,9	8,9	8,9	8,0	8,0	8,0	8,0	11,0	11,0	11,0	11,0
4b	13,5	13,5	13,5	13,5	20,0	20,0	20,0	20,0	18,0*	18,0*	18,0*	18,0*	8,9	8,9	8,9	8,9	8,0	8,0	8,0	8,0	11,0	11,0	11,0	11,0
4c	13,5	13,5	13,5	13,5	20,0	20,0	20,0	20,0	18,0*	18,0*	18,0*	18,0*	8,9	8,9	8,9	8,9	8,0	8,0	8,0	8,0	11,0	11,0	11,0	11,0
4d	13,5	13,5	13,5	13,5	20,0	20,0	20,0	20,0	18,0*	18,0*	18,0*	18,0*	8,9	8,9	8,9	8,9	8,0	8,0	8,0	8,0	11,0	11,0	11,0	11,0
5a	15,0*	15,0*	15,0*	15,0*	11,3	11,3	11,3	11,3	16,6	16,6	16,6	16,6	8,0	8,0	8,0	8,0	11,0	11,0	11,0	11,0	7,4	7,4	7,4	7,4
5b	15,0*	15,0*	15,0*	15,0*	11,3	11,3	11,3	11,3	16,6	16,6	16,6	16,6	8,0	8,0	8,0	8,0	11,0	11,0	11,0	11,0	7,4	7,4	7,4	7,4
5c	15,0*	15,0*	15,0*	15,0*	11,3	11,3	11,3	11,3	16,6	16,6	16,6	16,6	8,0	8,0	8,0	8,0	11,0	11,0	11,0	11,0	7,4	7,4	7,4	7,4
5d	15,0*	15,0*	15,0*	15,0*	11,3	11,3	11,3	11,3	16,6	16,6	16,6	16,6	8,0	8,0	8,0	8,0	11,0	11,0	11,0	11,0	7,4	7,4	7,4	7,4
6a	10,0	10,0	10,0	10,0	7,8	7,8	7,8	7,8	13,0*	13,0*	13,0*	13,0*	11,0	11,0	11,0	11,0	7,4	7,4	7,4	7,4	7,6	7,6	7,6	7,6
6b	10,0	10,0	10,0	10,0	7,8	7,8	7,8	7,8	13,0*	13,0*	13,0*	13,0*	11,0	11,0	11,0	11,0	7,4	7,4	7,4	7,4	7,6	7,6	7,6	7,6
6c	10,0	10,0	10,0	10,0	7,8	7,8	7,8	7,8	13,0*	13,0*	13,0*	13,0*	11,0	11,0	11,0	11,0	7,4	7,4	7,4	7,4	7,6	7,6	7,6	7,6
6d	10,0	10,0	10,0	10,0	7,8	7,8	7,8	7,8	13,0*	13,0*	13,0*	13,0*	11,0	11,0	11,0	11,0	7,4	7,4	7,4	7,4	7,6	7,6	7,6	7,6

Fonte: Elaboração das autoras.

iv) Servidores

O sistema do SAMU-Bauru/SP é composto por uma frota de nove ambulâncias distintas: oito USB´s e duas USA´s. Em 2013, as ambulâncias foram descentralizadas, localizadas em seis postos de saúde distribuídos um em cada setor da cidade, podendo deslocar-se para qualquer átomo para realizar um atendimento. As ambulâncias estão distribuídas de acordo com a Figura 4b.

v) Localização dos servidores

A matriz de localização (L) é facilmente obtida a partir da configuração original do sistema, de acordo com os critérios utilizados no SAMU-Bauru. Para generalizar a nomenclatura, a USA passa a ser chamado de veículo 1, enquanto as USB's passam a ser chamados de 2, 3, 4, 5, 6, 7, 8, 9 e 10, para os períodos da manhã e tarde. No período noturno há nove ambulâncias operando. A matriz de localização para o período da noite pode ser vista na Tabela 4.

Tabela 4 – Matriz de localização dos servidores nos átomos geográficos em 2013 no SAMU-Bauru.

	Nações	Geisel	Ipiranga	Bela Vista	Mary Dota	Boulevard			
	1	2	3	4	5	6			
	0	1	0	0	0	0	1	GA1	(VSA)
	0	1	0	0	0	0	2	GA2	(VSA)
	0	1	0	0	0	0	3	GB	(VSB)
	1	0	0	0	0	0	4	N	(VSB)
L=	0	0	1	0	0	0	5	IP	(VSB)
	0	0	0	1	0	0	6	BV1	(VSB)
	0	0	0	1	0	0	7	BV2	(VSB)
	0	0	0	0	1	0	8	MD	(VSB)
	0	0	0	0	0	1	9	BO	(VSB)

Fonte: Elaboração das autoras.

vi) Despacho dos servidores

O SAMU – Bauru/SP admite enviar apenas uma equipe por chamado, atendendo um dos requisitos para a aplicação do modelo hipercubo. As USAs atendem apenas aos chamados de emergência, caracterizando *backup* parcial. A formação de fila é permitida quando os usuários solicitam atendimento, enquanto todas as ambulâncias estão ocupadas. A fila de espera é formada e a escolha do próximo usuário a ser atendido é feita a partir da prioridade do chamado, na ordem do mais para o menos grave. Observa-se ainda que as USAs atendem apenas aos chamados vermelhos, enquanto que as USBs atendem a qualquer tipo de chamado.

vii) Política de despacho dos servidores

Ao receber um chamado e identificar a gravidade do caso, o médico regulador decide quem atenderá: uma USA ou USB. A política de despacho dos servidores depende da sua distribuição espacial e da localização deles. Lembrando que, na cidade de Bauru, os servidores estão descentralizados, portanto, a escolha do servidor preferencial é feita analisando a origem e a gravidade do chamado e a preferência é dada àquele localizado na mesma área (escolhido aleatoriamente). Se todos estiverem ocupados, é escolhido o primeiro servidor disponível mais próximo do chamado.

A lista de despacho em que um átomo tem mais de um servidor preferencial, ou seja, casos de desempate de prioridade entre ambulâncias do mesmo local, podem ser incorporadas no modelo hipercubo. Isso pode ser feito pela introdução da distribuição de frequências de despachos de cada servidor para cada átomo nas equações de balanço do sistema, ou considerando um número suficientemente grande de listas de preferências de despacho geradas aleatoriamente, o que representa, dessa forma, as possíveis chances dos servidores primários (e/ou *backup*) de cada átomo serem enviados para atender a um chamado em cada cenário investigado (BURWELL, et al. 1993; TAKEDA, 2000).

viii) Tempos de atendimento

Os tempos de atendimento são obtidos a partir do intervalo de tempo entre os instantes de saída e retorno à base. Esses valores estão disponíveis nas fichas de regulação médica. O tempo de atendimento é a soma dos tempos de preparo da equipe; tempo de viagem, definido como a saída da base até a chegada ao local; tempo em cena, desde o momento em que a ambulância chega ao local onde a equipe realiza o atendimento às vítimas, quando for o caso, até o momento em que ela sai do local; tempo de viagem de volta, desde o instante de saída do local até o momento em que a ambulância retorna à base. Para verificar estatisticamente a hipótese de que os tempos de atendimento são exponencialmente distribuídos, foi usado o teste de Kolmogorov-Smirnov. Em todas as ambulâncias, foi rejeitada a hipótese de que os tempos de serviço são exponencialmente distribuídos, com $\alpha = 5\%$ de significância.

A Tabela 5 mostra os tempos médios de atendimento (em minutos), desvios-padrão, coeficiente de variação e as taxas médias para cada ambulância. Para verificar a hipótese de diferenças nos tempos de atendimento entre os servidores, foi realizada a análise de variância ANOVA (COSTA NETO, 1977; MAGALHÃES; LIMA, 2002) com nível de significância α = 5%. Os resultados mostraram que as diferenças entre as médias dos tempos de atendimento entre os servidores são significativas nos três períodos. Dessa forma, a aplicação do modelo hipercubo deve considerar que os servidores não são homogêneos para os três períodos considerados. De acordo com Larson (1974, 2007) e Jarvis (1985), esse tipo de sistema pode ser analisado aproximadamente pelo modelo hipercubo sem que a análise seja muito comprometida. Essa aproximação também foi feita em Takeda (2007), Iannoni et al. (2006) e Souza (2010) sem que a análise dos sistemas fosse comprometida.

O tempo médio de atendimento é diferente do desvio-padrão, ou seja, o coeficiente de variação é bem menor que 1, indicando que os tempos de atendimento não são exponencialmente distribuídos.

Tabela 5 – Tempos e taxas médias de atendimento para cada ambulância.

Ambulância	Tempo médio de atendimento (minutos)	Desvio-padrão (*)	Coeficiente de variação	μ (horas)
1 - GA1	55	12	0,2	1,0856
2 - GA2	55	16	0,3	1,0909
3 - GB	45	15	0,3	1,3245
4 - N	40	14	0,3	1,4957
5 - IP	41	11	0,3	1,4683
6 - BV1	48	19	0,4	1,2429
7 - BV2	43	16	0,4	1,3907
8 - MD	46	19	0,4	1,3034
9 - BO	43	15	0,3	1,3863
USA	55	14	0,3	1,0883
USB	44	15	0,3	1,3731
Total	46	15	0,3	1,3098

Fonte: Elaboração das autoras.

ix) Relação entre o tempo de atendimento e o tempo de viagem

É necessário verificar se os tempos médios de viagem são pequenos em relação aos tempos médios de atendimento para cada ambulância. A Tabela 6 mostra o tempo médio de serviço, o tempo médio de viagem e a relação entre o tempo médio de atendimento e o tempo médio de viagem para cada servidor. Pode-se notar que os tempos médios de viagem são relativamente pequenos com relação aos tempos médios de atendimento. Os tempos médios de viagem representam, no máximo, 40% do tempo total de atendimento das ambulâncias 2 e 3, no período da noite, de forma que a hipótese 9 do modelo hipercubo está validada.

Tabela 6 – Relação entre o tempo de atendimento e o tempo de viagem para as ambulâncias.

Ambulâncias		Tempo médio de viagem	Tempo médio de atendimento (minutos)	Relação: Tempo médio de viagem/ Tempo médio de atendimento
1 -	GA1	13,6	55	0,2473
2 -	GA2	11,5	55	0,2091
3 -	GB	11,3	45	0,2511
4 -	N	9,3	40	0,2325
5 -	IP	10,1	41	0,2463
6 -	BV1	12,5	48	0,2604
7 -	BV2	9,0	43	0,2093
8 -	MD	10,2	46	0,2217
9 -	BO	7,2	43	0,1674
USB		9,9	55	0,2270
USA		12,6	44	0,2282
Sistema		10,5	46	0,2272

Fonte: Elaboração das autoras.

5. APLICAÇÃO DO MODELO: COMPARAÇÃO MODELO ORIGINAL VS AMOSTRA

A partir da resolução do Sistema 2 foi possível calcular as medidas de desempenho apresentadas na Seção 2.3 e realizar a comparação dos resultados obtidos com as informações retiradas da amostra. A Tabela 7 mostra a carga de trabalho (*Workload*) dos servidores no sistema modelado em comparação aos resultados da amostra e seus desvios relativos. Nota-se uma melhor aderência para os servidores onde a carga de trabalho foi distribuída, como os servidores avançados (GA1 e GA2) e os servidores da região Bela Vista (BV1 e BV2). Contudo, para os outros servidores vê-se desvios relativos de maior ordem, chegando a uma diferença de 32%.

Tabela 7 – Comparação das cargas de trabalho dos servidores.

	Carga de Trabalho		
Servidores	Amostra	Modelo	Desvio-relativo (%)
GA1	0,1619	0,1622	0,16
GA2	0,1619	0,1587	2,00
GB	0,3706	0,2512	32,21
NÇ	0,3121	0,3241	3,85
IP	0,2497	0,2948	18,05
MD	0,3792	0,3214	15,24
BV1	0,3500	0,3654	4,40
BV2	0,3500	0,3669	4,83
BLV	0,3008	0,3862	28,38

Fonte: Elaboração das autoras.

A Tabela 8 compara os tempos médios de espera em fila para cada uma das prioridades entre os resultados do modelo hipercubo e da amostra coletada. Os desvio-relativos absolutos também estão contidos na tabela. Aqui fica bastante clara a interferência do fator humano para o despacho das equipes, o modelo presume que se algum servidor que possa atender a um chamado esteja livre ele será enviado imediatamente, contudo o sistema, em especial para baixas prioridades, busca esperar a chegada de chamados de maior gravidade para que esses sejam melhor atendidos.

Tabela 8 – Comparação dos tempos médios de espera em fila.

	wq (h)		
Prioridades	Amostra	Modelo	Desvio-relativo (%)
a	0,2341	0,0093	96,03
b	0,1580	0,0110	93,04
c	0,3745	0,0144	96,15
d	0,5728	0,0173	96,98
Geral	0,2896	0,0127	95,61

Fonte: Elaboração das autoras.

Os tempos médios de viagem para cada subátomo são mostrados na Tabela 9. Sendo que, neste caso, computou-se também o tamanho da amostra a fim de se verificar a sua influência na comparação com o modelo hipercubo. Novamente se observa a falta de aderência dos dados apresentados pela amostra em comparação aos resultados do modelo. Contudo, vale ressaltar o problema em se conseguir dados realmente confiáveis da amostra, visto que esta precisa de um tamanho adequado, logo se nota que em praticamente metade dos átomos (11 de 24) possuem uma amostra de chamados iguais ou menores do que 5.

Tabela 9 – Comparação dos tempos médio de viagem para cada subátomo.

Ta					
Átomos		Tamanho amostra	Amostra	Modelo	Desvio-relativo (%)
1 -	1d	1	10,0	14,1	41,33
2 -	1c	12	12,5	14,1	12,84
3 -	1b	15	9,3	14,9	60,28
4 -	1a	5	9,6	13,8	43,88
5 -	2d	5	17,6	9,4	46,80
6 -	2c	20	7,7	10,1	31,56
7 -	2b	23	7,5	9,5	26,55
8 -	2a	3	12,7	13,5	6,89
9 -	3d	3	12,3	10,9	11,71
10 -	3c	11	12,2	10,7	11,94
11 -	3b	9	6,6	10,7	63,61
12 -	3a	4	13,3	15,8	19,09
13 -	4d	2	13,5	10,0	26,07
14 -	4c	15	9,3	10,0	7,98
15 -	4b	7	5,3	11,5	117,62
16 -	4a	2	10,5	13,4	27,95
17 -	5d	7	11,1	11,7	5,38
18 -	5c	18	8,7	11,5	32,19
19 -	5b	23	7,9	11,7	49,22
20 -	5a	6	17,0	14,9	12,56
21 -	6d	1	18,0	9,2	48,67
22 -	6c	3	14,0	8,2	41,40
23 -	6b	6	6,0	8,3	38,71
24 -	6a	1	14,0	9,9	29,05

Fonte: Elaboração das autoras.

A Tabela 10, mostra a comparação desses tempos em relação a amostra e ao modelo. Nota-se aqui que todos servidores possuem mais de 10 chamados atendidos na amostra para a obtenção dos dados. Contudo, vale ressaltar a aderência para os servidores avançados, com desvios menores de 10%. Por outro lado, para os servidores básicos, não houve uma boa aderência, com desvios relativos chegando a 30,49%.

Tabela 10 – Comparação dos tempos médios de viagem dos servidores.

		Tu		
Servidores	Tamanho amostra	Amostra	Modelo	Desvio-relativo (%)
GA1	11	13,6	14,5	6,71
GA2	11	13,6	14,2	4,68
GB	29	10,8	14,0	30,51
NÇ	21	8,4	9,8	17,05
IP	22	8,8	11,4	30,49
MD	30	10,0	10,8	8,31
BV1	27	9,0	11,8	30,96
BV2	27	9,0	11,5	28,29
BLV	25	7,3	9,4	27,95

6. CONCLUSÃO

Dada a relevância dos Sistemas de Atendimento Emergenciais (SAE's) para a sociedade, principalmente no que se refere aos tempos de resposta desses serviços. Um fator importante a ser considerado é que devido às restrições orçamentárias, os SAE's não podem ter um grande número de pessoas e equipamentos. Gerando, dessa forma, um compromisso evidente entre investimentos, custos operacionais e o nível de serviço oferecido aos usuários.

A fim de dar subsídios para tomada de decisão, uma forma objetiva de avaliar o sistema principalmente em períodos de alta demanda é importante para os gerentes do sistema. Nesse contexto, os objetivos desse trabalho foram: (i) descrever os chamados e os atendimentos do SAMU-Bauru/SP, (ii) aplicar o modelo hipercubo e (iii) obter suas principais medidas de desempenho. Para isso, realizou-se um estudo de caso no Serviço de Atendimento Móvel de Urgência (SAMU) no município de Bauru.

Os chamados e os atendimentos do SAMU/Bauru foram coletados durante o ano de 2013 e organizados em tabelas, gráficos e medidas descritivas de forma a obter o período de pido do sistema analisado, que foi das 12h às 18h. Nesse período foram coletados todos os dados necessários para o estudo, conforme descrito na Seção 3.

Para aplicação do modelo hipercubo é necessária a verificação das nove hipóteses descritas na Seção 4. Todas foram verificadas, com exceção da hipótese 8, dos tempos de serviço serem exponencialmente distribuídos. Apesar dessa hipótese ter sido rejeitada, o consideramos que essa seria uma aproximação razoável, dado que estudos similares levantados na literatura mostraram que mesmo com essa hipótese sendo rejetada o modelo ainda foi uma boa aproximação para os estudos como, por exemplo, em Takeda et al. (2007) e Iannoni et al. (2006).

O modelo hipercubo tem por objetivo avaliar a configuração e estimar as medidas de desempenho para SAE's, possibilitando um planejamento adequado e melhores níveis de serviço oferecido. Dessa forma foram calculadas as cargas de trabalho dos servidores, os VSA's tiveram sua carga de trabalho entre 0,15 e 0,16 enquanto que os VSB's ficaram entre 0,24 e 0,38. Todos os resultados foram comparados com a amostra. Os tempos médios de viagem, de uma forma geral, ficaram entre 7,3 minutos e 15,8 minutos. Os tempos médios de espera em fila para cada uma das prioridades tiveram uma diferença grande do modelo em relação a amostra, da ordem de 95%.

Os resultados mostram desvios relativos altos principalmente em chamados básicos, menos urgentes, isso parece indicar a interferência do fator humano para o despacho das equipes. O modelo presume que se algum servidor que possa atender a um chamado esteja livre ele será enviado imediatamente. Contudo, o sistema, em especial para baixas prioridades, busca esperar a chegada de chamados de maior gravidade para o atendimento desses chamados. Isso caracteriza um sistema com reserva de servidores.

Uma perspectiva futura a ser considerada é a análise transiente no SAMU/Bauru a fim de verificar o impacto da abordagem clássica do modelo hipercubo e da abordagem transiente quando somente uma parte do sistema entra em equilíbrio.

Agradecimento: Os autores agradecem ao SAMU – Bauru/SP pela colaboração com essa pesquisa, à Fapesp e ao CNPq.

REFERÊNCIAS

ALANIS, R.; INGOLFSSON, A.; KOLFAL, B. A Markov chain model for an EMS system with repositioning. **Production and operations management**, v. 22, n. 1, p. 216-231, 2013.

BRANDEAU, M. L.; LARSON, R. C. Extending and applying the hypercube queueing model to deploy ambulances in Boston. National Emergency Training Center. **TIMS Studies in the Management Science**, v. 22, p. 121-53, 1986.

BURWELL, T. H.; JARVIS, J. P.; MCKNEW, M. A. Modeling co-located servers and dispatch ties in the hypercube model. **Computers & Operations Research**, v. 20, n. 2, p. 113-119, 1993.

CHELST K. R.; BARLACH Z. Multiple unit dispatches in emergency services: models to estimate system performance. **Management Science**, v. 27, n. 12, p. 1390-1409, 1981.

CHIYOSHI F.; GALVÃO R. D.; MORABITO R. O uso do modelo hipercubo na solução de problemas de localização probabilísticos. **Gestão & Produção**, v. 7, n. 2, p. 146-174, 2000.

COSTA D. M. **Uma metodologia iterativa para determinação de zonas de atendimento de serviços emergenciai**s. 2004. 132 f. Tese (Doutorado em Engenharia de Produção) - Universidade Federal de Santa Catarina, Santa Catarina, 2004.

FIGUEIREDO A. P. S.; LORENA L. A. N. Localização de ambulâncias: uma aplicação para a cidade de São José dos Campos. *In*: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 21, 2005, Goiânia, GO, **Anais...** Goiânia GO: INPE, 2005.

GALVÃO R. D.; MORABITO, R. Emergency service systems: The use of the hypercube queuing model in the solution of probabilistic location problems. **International Transactions in Operational Research**, v. 15, n. 5, p. 1-25, 2008.

GONÇALVES M. B.; NOVAES A. G.; ALBINO J. C. C. Modelos para localização de serviços emergenciais em rodovias. In: SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL, 26, 1994, Florianópolis, SC. **Anais...** Florianópolis SC: SPO, 1994.

GONÇALVES M. B.; NOVAES A. G.; SCHMITZ R. Um modelo de otimização para localizar unidades de serviço emergenciais em rodovias. *In*: CONGRESSO DE PESQUISA E ENSINO EM TRANSPORTES, 9, 1995, São Carlos, SP. **Anais...** São Carlos SP: CPET, 1995.

IANNONI A. P.; MORABITO R. Modelo hipercubo integrado a um algoritmo genético para análise de sistemas médicos emergenciais em rodovias. **Gestão & Produção**, v. 13, n. 1, p. 93-104, 2006.

LARSON R. C. Hypercube queuing model for facility location and redistricting in urban emergency services. **Computers and Operations Research**, v. 1, n. 1, p.67-95, 1974.

LARSON R. C.; ODONI A. R. **Urban Operations Research**. Dynamic Ideas, Belmont, Massachusetts, 2007.

LITTLE J. D. A proof for the queuing formula. **Operations Research**, v. 9, p.383-387, 1961.

LOPES, S. L.B.; FERNANDES, R. J. **Uma breve revisão do atendimento médico pré-hospitalar**. Medicina (Ribeirao Preto. Online), v. 32, n. 4, p. 381-387, 1999.

LUQUE, L. **Análise da aglutinação de estados em cadeias de markov do modelo hipercubo de filas com servidores co-localizados**. 2006. Dissertação (Mestrado) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos, SP, 2006.

MENDONÇA F.; MORABITO R. Analyzing emergency medical service ambulance deployment on a Brazilian highway using the hypercube model. **Journal of the Operational Research Society**, v. 52, n. 3, p. 261-270, 2001.

OLIVEIRA L. K. **Uma aplicação do modelo hipercubo de filas para avaliação do centro de emergência da polícia militar de Santa Catarina**. 2003. 92 f. Dissertação (Mestrado em Engenharia de Produção) - Universidade Federal de Santa Catarina, Florianópolis, SC, 2003.

SACKS S. R.; GRIEF S. Orlando Police Department uses OR/MS methodology, new software to design patrol districts. **OS/MS Today**, Baltimore, 1994.

SOUZA, R. M. ; MORABITO, R ; CHYIOSHI, Y. F. ; IANONNI, A. P. Análise da configuração de SAMU utilizando múltiplas alternativas de localização de ambulâncias. **Gestão & Produção**, v. 20, p. 287-302, 2013.

SOUZA R. M. ; MORABITO R..; CHIYOSHI, F. Y.; IANONNI A. P. 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**, v. 242, n. 1, p. 274-285, 2015.

SWERSEY A. J. **Handbooks in OR/MS**. Amsterdam: Elsevier Science, v. 6, p. 151-200, 1994.

TAKEDA, R. A. **Uma contribuição para avaliar o desempenho de sistemas de transporte emergencial de saúde**. 2000. 210 f. Tese (Doutorado em Transportes) - Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, SP, 2000.

TAKEDA R. A.; WIDMER J. A.; MORABITO R. Analysis of ambulance decentralization in an urban emergency medical service using the hypercube queueing model. **Computers & Operations Research**, v. 34, n. 3, p. 727-741, 2007.