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

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

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

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

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.

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.*

1 Introdução

Só no ano de 2014 ocorreram cerca de 140 mil acidentes de trânsito no Brasil (DNIT, 2015). Embora esse dato evidencie redução, em relação aos anos anteriores, o Brasil ainda é o terceiro país com mais mortes causadas pelo trânsito no mundo (OMS, 2016). Isso faz com que o atendimento pré-hospitalar (um dos menores componentes de custos em acidentes) tenha uma importância fundamental na redução de mortes e lesões severas causadas por acidentes de trânsito (Lopes & Fernandes, 1999).

O atendimento pré-hospitalar, como explicado, é parte de um Sistema de Atendimento Emergencial (SAE). Esses sistemas são estudados desde os anos 1950 por meio da Pesquisa Operacional e têm

[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, Brasil, e-mail: beojone@hotmail.com; regiane@feb.unesp.br

Recebido em Jun. 13, 2016 - Aceito em Nov. 3, 2016

Suporte financeiro: CAPES e FAPESP.

mostrado potencial de melhoria em bombeiros, polícia, Sistema de Atendimento Móvel de Urgência (SAMU), dentre outros. A ideia é encontrar meios de fornecer serviços de saúde efetivos e melhorar a qualidade de vida da população considerando o *trade-off* existente entre nível de serviço e a escassez de recursos (Simpson & Hancock, 2009; Souza, 2010).

O SAMU é um sistema médico emergencial público oriundo de um programa do governo federal realizado a partir de um acordo bilateral entre Brasil e França. O sistema brasileiro é baseado no modelo francês, que já opera há mais de 30 anos (Takeda et al., 2004). O SAMU opera 24 horas por dia, 7 dias por semana e conta com equipes médicas compostas por médicos, enfermeiros e auxiliares em enfermagem. Os chamados são realizados através do número de telefone 192 e são classificados conforme sua localização e nível de urgência (gravidade), para que seja enviada uma equipe em uma ambulância, que pode atender em vias públicas, locais de trabalho e em residências (Ghussn & Souza, 2016).

SAE como os SAMU são caracterizados essencialmente por incertezas principalmente quanto a disponibilidade, localização, tempo de serviço dos servidores, demanda ao longo da região e tempo de resposta aos usuários. Os SAE em saúde, em geral, caracterizam um grande desafio para o sistema de saúde. Independentemente do tipo de urgência envolvida, somente com uma rigorosa organização é possível oferecer um serviço de boa qualidade (Souza, 2010).

O modelo hipercubo de filas é um modelo descritivo, baseado na teoria das filas, que possibilita calcular medidas de desempenho relevantes para um sistema de atendimento de urgência, que 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 (átomo geográfico) 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 (Larson & Odoni, 2007).

O modelo hipercubo de filas tem sido estudado em diversos trabalhos desenvolvidos no Brasil, como 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). Em outras partes do mundo há aplicações em Chelst & Barlach (1981), Brandeau & Larson (1986), Burwell et al. (1993), Sacks & Grief (1994). Swersey (1994) e Larson & Odoni (2007). Em todos os estudos, o modelo hipercubo se mostrou eficiente e preciso.

Nesse contexto, o objetivo do presente trabalho é 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 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.

Uma descrição do sistema SAMU de Bauru é apresentada na seção 2, assim como a explicação do seu processo de atendimento. A seção 3 traz uma explicação sobre o modelo hipercubo utilizado, com suas particularidades. Na seção 4 mostram-se os testes das hipóteses para aplicação do modelo. Na seção 5 tem-se a descrição de como foi feita a aplicação do modelo, tanto para o cenário original quanto para os alternativos. A análise dos resultados foi feita na seção 6. Por fim, a seção 7 apresenta as considerações finais dos autores e perspectivas de pesquisas futuras.

2 O SAMU Bauru

Segundo JCNet (2010), em 2010, o SAMU da cidade de Bauru integrava 17 cidades da região numa parceria com prefeituras. A sede de atendimento está em Bauru, que deve receber os chamados e distribuí-los de acordo com a gravidade dos casos e cada região. Eram aproximadamente 90 profissionais envolvidos, dentre eles 35 condutores e 32 auxiliares, o serviço contava com 25 ambulâncias, entre avançadas e básicas. O serviço regionalizado é responsável pela população de Bauru, aproximadamente 400 mil habitantes, 800 mil pessoas, na região como um todo.

A coleta dos dados utilizou relatórios fornecidos pelo próprio SAMU referentes a 2012 e 2013. Os dados coletados referem-se a 10 dias do mês de setembro de cada ano. Um resumo da coleta de dados e verificação das hipóteses está na seção 4.

O processo de atendimento de um chamado em um SAMU pode ser descrito conforme a Figura 1. Para começar o processo é preciso que haja um chamado, normalmente realizado pelo telefone 192. Caso haja uma ambulância livre, a equipe para o atendimento é enviada e a ambulância sai da base assim que feita a classificação (triagem) do chamado. Após isso, começa a viagem até o local do chamado. O intervalo de tempo entre o recebimento do chamado e a chegada ao local de atendimento representa o tempo de resposta. Começa então o atendimento ao usuário em si, o qual pode variar conforme os cuidados necessários. O paciente é então levado para um destino, o qual pode ser hospital ou pronto-atendimento. Após deixar o paciente em seu destino, a ambulância retorna à base, o que marca o final do serviço (Schmid, 2012).

O serviço do SAMU envolve ainda a classificação dos chamados conforme sua gravidade. A classificação é feita utilizando-se um esquema de cores: vermelho

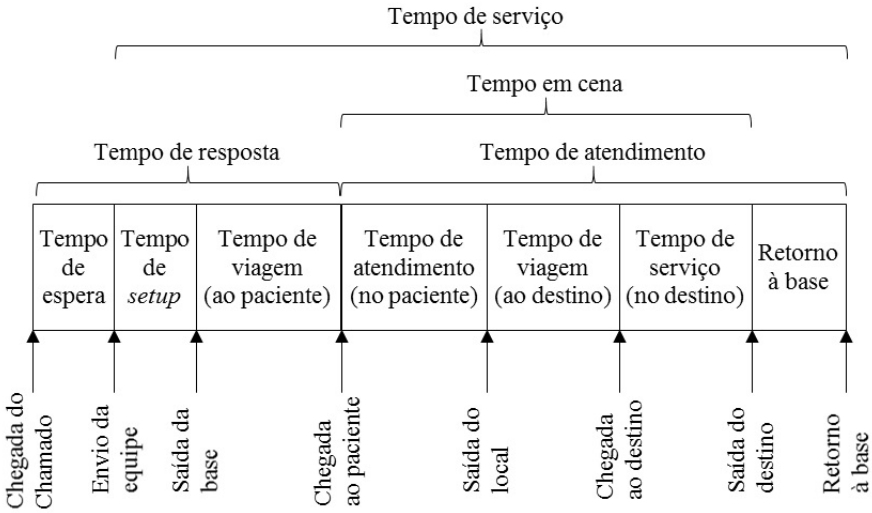

Figura 1. Linha do tempo com os eventos do atendimento de uma ambulância.

(mais grave), amarelo, verde e azul (menos grave). O médico realiza a classificação dos chamados conforme as informações do solicitante no momento da ligação. Essa classificação também é utilizada para decidir sobre a política de despacho das ambulâncias, já que os VSA (Veículos de Serviço Avançado) são enviados apenas aos chamados classificados como vermelho, que incorrem em risco de vida ao paciente. Essa política de despacho, chamamos de Backup Parcial pois os VSA não são enviados para atender os chamados menos graves (amarelo, verde e azul). Em particular, o SAMU de Bauru tem dois servidores VSA, diferentemente do universo de estudo de Souza et al. (2014), que tinha apenas um servidor preferencial no sistema.

3 Modelo hipercubo

Desenvolvido por Larson (1974), o modelo hipercubo de filas é um modelo analítico, baseado em sistemas de filas espacialmente distribuídas em que a ideia é fazer uma expansão dos estados de um sistema *M/M/m* buscando-se representar os *m* servidores (indistinguíveis) individualmente em um sistema em que o servidor vai até o usuário. Dessa maneira, ele permite trabalhar com políticas de envio de ambulâncias mais complicadas. Para encontrar a solução do sistema é preciso elaborar e resolver um conjunto de equações de equilíbrio (*steady state*), os resultados são as probabilidades de ocorrência de estado de equilíbrio. A partir da solução do modelo podem ser calculadas diversas medidas de desempenho para o sistema, como: carga de trabalho dos servidores, frequências de despacho de servidores, tempos médios de viagem, tempos médios de fila, entre outras (Souza, 2010).

Para a utilização em diversas abordagens foi necessário criar adaptações do modelo clássico proposto por Larson (1974), de forma a se trabalhar com um modelo mais próximo à realidade do sistema analisado. Essas modificações são denominadas extensões do modelo. Entre as extensões vale citar algumas, presentes no SAMU Bauru, como: prioridade em fila e aleatoriedade de despacho.

O uso do modelo hipercubo envolve o uso de vários parâmetros. O uso de extensões do modelo pode fazer com que alguns desses parâmetros seja alterado. Na Tabela 1 são apresentadas as notações para esses parâmetros. Os conceitos sobre subátomos e prioridades são apresentados na seção 3.3.

No hipercubo, o espaço de estados indica a disponibilidade de cada servidor individualmente, conforme mencionado anteriormente. Considere um sistema com $m = 3$ servidores, há $2^3 = 8$ possíveis estados do sistema: {000}, {001}, {010}, ..., {111}. Os 0 e 1 indicam se cada um dos três servidores está livre ou ocupado, respectivamente. O espaço de estados de um sistema com três servidores é representado por um cubo, no caso de haver mais de três servidores, temos um hipercubo.

A Figura 2 ilustra o espaço de estados desse sistema com três servidores. É possível tratar um sistema em que é permitida ou não a formação fila no modelo hipercubo; no modelo clássico, os chamados que chegam esperam em uma fila por meio da qual os usuários são atendidos a partir do primeiro servidor a se desocupar, segundo a disciplina FCFS. Os estados S4, S5, S6, ... da Figura 2 representam os estados com 1, 2, 3, ... usuários na fila de espera do sistema, respectivamente.

Larson & Odoni (2007) mostram uma série de nove hipóteses que precisam ser satisfeitas para a aplicação do modelo hipercubo em sua forma clássica:

1. Existência de átomos geográficos: A região onde são prestados os serviços do sistema

Tabela 1. Notações utilizadas.

Notação	Significado	Unidade de medida
N_A	Número de átomos	Número
N	Número de servidores	Número
τ_{ij}	Tempo de viagem do átomo *i* ao átomo *j*	Minutos
$\tau_{ik,jh}$	Tempo de viagem do subátomo *ik* ao subátomo *jh*	Minutos
λ_j	Taxa de chegada de chamados do átomo *j* $\left(\lambda_j = \sum_k \lambda_{jk}\right)$	Chamados/hora
λ_{jk}	Taxa de chegada de chamados do subátomo *jk*	Chamados/hora
λ_k	Taxa de chegada de chamados de prioridade *k*	Chamados/hora
λ	Taxa total de chegada do sistema	Chamados/hora
μ_i	Taxa de serviço do servidor *i*	Chamados/hora
μ	Taxa total de serviço do sistema	Chamados/hora
μ^{-1}	Tempo de serviço	Minutos

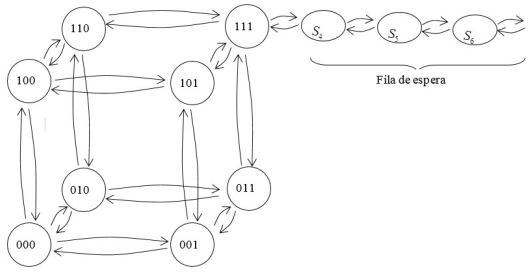

Figura 2. Estados do modelo hipercubo com três servidores.

deve ser dividida em N_A átomos geográficos, sendo que cada átomo corresponde a uma fonte independente de chamados e também possui políticas de despacho.

2. Processo de chegada conforme a distribuição de Poisson: Os usuários de cada átomo solicitam chamados por meio do processo de Poisson, sendo os chamados independentes entre si. Além disso, as taxas de chegada, λ_j, de chamados de cada átomo deve ser conhecida.

3. Tempo de viagem dos servidores: Os tempos de viagem, τ_{ij}, de cada servidor *i* para o átomo *j* devem ser conhecidos ou estimados.

4. Servidores: Existem *N* servidores espacialmente distribuídos ao longo do sistema, sendo que cada um pode se deslocar e atender a qualquer um dos átomos.

5. Localização do servidor: A localização do servidor no sistema deve ser conhecida ao menos probabilisticamente, sendo que o servidor pode se mover pelos átomos ou ficar fixo em um deles.

6. Despacho simples: Para atender qualquer chamado é enviado apenas um servidor para o local. Se não houver servidores disponíveis, os chamados entrarão em fila ou serão considerados perdas do sistema.

7. Política de despacho dos servidores: Há uma lista (matriz) de preferência de despacho para cada átomo, ou seja, deve ser obedecida uma ordem de envio dos servidores para os chamados, a qual deve ser bem estabelecida. Por exemplo, em um sistema com 4 servidores e 3 átomos, ao chegar um chamado do átomo 3, deve-se obedecer a lista de preferência de (1, 3, 4, 2), ou seja, o servidor 1 é o primeiro a ser enviado, caso esteja ocupado envia-se o servidor 3 e assim por diante, até o servidor 2.

8. Tempo de serviço: O tempo de serviço de um servidor engloba tempo de *setup*, tempo de viagem, tempo em cena até o retorno à base.

9. Dependência do tempo de serviço em relação ao tempo de viagem: A variação do tempo de viagem deve ser considerada uma variável de segunda ordem no tempo total de serviço, quando comparado ao tempo em cena e em preparação da equipe. Isso não quer dizer que o tempo de viagem seja ignorado no cálculo do tempo médio de serviço, já que esse é incorporado através da calibração do tempo médio de serviço (μ^{-1}), em que μ^{-1} é igual à soma do tempo médio de viagem do servidor e do tempo médio de atendimento.

As extensões do modelo clássico trabalham com alterações nessas hipóteses, de acordo com a necessidade do sistema estudado. Exemplos de aplicação do modelo em sua forma original podem ser facilmente encontrados na literatura para a melhor compreensão do funcionamento do modelo. A seguir é apresentado um exemplo de aplicação do modelo

hipercubo clássico encontrado em Chiyoshi et al. (2000).

Considere um sistema, que não admite fila de espera, cuja região é dividida em três átomos, que são atendidos conforme a matriz de preferência de despacho na Tabela 2.

A construção das equações de equilíbrio mostra a solução para o modelo em equilíbrio (Equações 1-8). Note-se que, devido à matriz de preferência de despacho, a transição do estado {100} para o estado {110} possui uma taxa de $\lambda_1 + \lambda_2$, já que o servidor preferencial do átomo 1, servidor 1, está ocupado e a primeira opção para envio é o servidor 2, somando-se a taxa de chegada do átomo 2, em que o servidor 2 é o preferencial. Esse sistema pode ser resolvido como um sistema linear homogêneo determinado caso uma de suas equações de equilíbrio seja substituída pela equação $\sum_{B\in D} P_B = 1$, que mostra que a soma das probabilidades do sistema é igual à 1.

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

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

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

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

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

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

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

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

A partir das probabilidades de estado pode-se calcular diversas medidas de desempenho para o sistema.

A carga de trabalho (*workload*) de um servidor do sistema é a fração de tempo em que o servidor está ocupado em média. Para calcular essa importante medida de desempenho é necessário somar as probabilidades de o servidor estar ocupado, as probabilidades de estado em que estão em serviço {1}, mais as probabilidades dos estados em fila.

Tabela 2. Matriz de preferência de despacho para o exemplo.

Átomo	Preferência		
	1º	2º	3º
1	1	2	3
2	2	3	1
3	3	1	2

Outra medida de desempenho importante é a frequência de despacho. O seu cálculo envolve os despachos realizados sem espera (*nq*) e com espera (*q*), conforme mostra a Equação 9.

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

em que λ_j / λ é a fração das chegadas correspondentes ao átomo *j* no sistema; $\sum_{B\in E_{nj}} P_B$ é a soma das probabilidades de envio do servidor *i* ao átomo *j*, quando o servidor *i* estiver livre e for o próximo a ser enviado ao átomo *j* pela matriz de preferência de despacho; P_Q' é a probabilidade de saturação do sistema, a soma da probabilidade dos estados de fila mais a probabilidade de todos servidores estarem ocupados; μ_i / μ é a probabilidade de o servidor *i* ser o primeiro liberado, quando todos servidores estiverem ocupados.

O tempo médio que um servidor *i* leva para viajar ao átomo *j*, quando disponível, calcula-se conforme a Equação 10.

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

O tempo médio de viagem para chamados sujeitos à espera em fila é calculado conforme na Equação 11.

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

O tempo médio de viagem do sistema pode ser calculado pela Equação 12.

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

O tempo médio de viagem ao átomo *j* é dado pela Equação 13.

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

Os tempos médios de viagem do servidor *i* são dadso pela Equação 14.

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

O tempo de espera do sistema pode ser calculado por meio da Fórmula de Little, conforme indicado na Equação 15 (Little, 2011).

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

3.1 Prioridade em fila

Em sistemas que consideram prioridade em fila, classes de usuários são definidas em termos de suas prioridades. Para fins de modelagem, os átomos geográficos são divididos em camadas (*layers*), de forma que cada camada representa uma prioridade específica do sistema (Takeda et al., 2007). A Figura 3 ilustra o processo de separação dos chamados em camadas, conforme sua prioridade em um sistema com três prioridades: *a*, *b* e *c*. Essas camadas formam subátomos não são físicos dos átomos.

A prioridade envolve a ordem pela qual os chamados são atendidos na fila, de acordo com a gravidade do chamado, ou seja, sua classificação. (Souza et al., 2015). A Figura 4 mostra um exemplo, apresentado em Souza et al. (2015), em que ocorrem mudanças de estado de fila onde há prioridade. Os nós representam os estados (situação) do sistema, por exemplo, o estado $\{abb\}$ é a situação em que há um chamado de prioridade *a* (alta prioridade) e dois de prioridade *b* (média prioridade) na fila de espera. Os arcos que ligam os estados representam as transições entre os estados, por exemplo, o sistema sai da situação $\{ab\}$ e entra na situação $\{b\}$ quando um servidor termina seu serviço e começa a atender o primeiro chamado em fila. Note que a classe de chamados *a* tem prioridade em relação à classe de chamados *b*, assim como a classe *b* tem prioridade em relação à classe *c*. Enquanto houver um chamado *a* em fila, o chamado *b* não será atendido. Exemplos ilustrativos podem ser encontrados em Souza (2010).

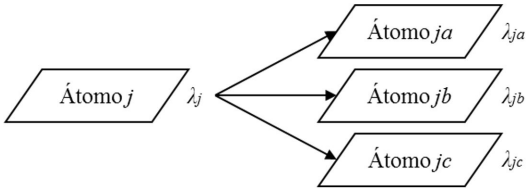

Figura 3. Ilustração do *layering*.

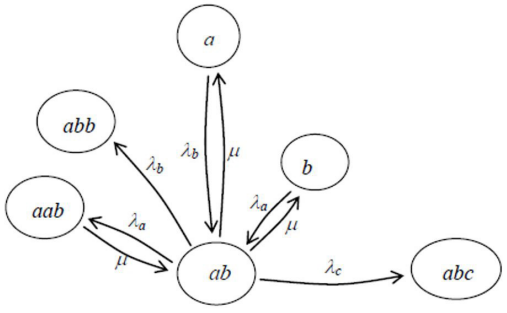

Figura 4. Mudanças de estado em fila com prioridade. Fonte: Souza et al. (2015, p. 276).

A partir dessas transições de estado podem ser construídas equações de equilíbrio para os estados de fila com prioridade conforme a Equação 16 para o estado $\{ab\}$, mostrado na Figura 4.

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

A principal contribuição nas medidas de desempenho é a possibilidade de se calcular o tempo de espera na fila para cada tipo de chamado (Souza et al., 2015). Se definirmos n_r como o número classes de *r* usuários na fila, a probabilidade $P(n_r = j)$ é dada pela soma das probabilidades associadas com os estados da fila (Equação 17):

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

em que *S* é o estado da fila e $n(r,S)$ é o número da classe de *r* usuários em *S*. Dessa distribuição, o número médio de *k* usuários pode ser determinado pela Equação 18:

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

O tempo médio de espera na fila pode ser obtido a partir da Fómula de Little (Equação 19):

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

3.2 Aleatoriedade de despacho

Esta extensão busca representar sistemas onde não há uma matriz de preferência de despacho bem definida, casos nos quais certa vez um servidor teve a preferência para atender a um chamado de um átomo, enquanto em outro momento poderia ser outro o servidor enviado, não havendo regra bem definida para o envio do servidor.

Existem ao menos duas maneiras de se representar um sistema assim. Primeiramente, pode-se construir e resolver vários modelos com uma lista de preferência de despacho definida a cada resolução do modelo, assim como em Takeda et al. (2004). Ao final das resoluções, chega-se a uma média dos resultados para encontrar as probabilidades finais de cada estado. Outra forma, conforme Chiyoshi et al. (2011), é construir e resolver apenas um sistema em que a taxa de transição de um estado, por exemplo, $\{000\}$, para um estado $\{001\}$, em que todos os servidores estão no mesmo local e não há preferência de envio para nenhum deles, é de 1/3 de toda a taxa de chegada do sistema. Nesse caso, a taxa de transição de estado é distribuída entre os servidores livres para atendimento, como mostra a Equação 20.

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

Larson (1974) e Batta et al. (1989) mostraram uma forma que utiliza fatores de correção para escolher os servidores enviados. A ideia é que seja enviado o servidor mais próximo do chamado e a localização dos servidores é conhecida probabilisticamente. A probabilidade de um servidor ser enviado é proporcional ao produto da probabilidade desse servidor estar disponível e da probabilidade de o servidor preferencial estar ocupado. Esse processo não é exatamente aleatório, visto que os fatores de correção são na realidade determinísticos.

4 Verificação das hipóteses para aplicação do modelo hipercubo

4.1 Área dividida em N_A átomos geográficos

No ano de 2013, o SAMU Bauru separava seus chamados de acordo com a área de origem, átomos geográficos. O mapa, com a divisão das áreas, está na Figura 5.

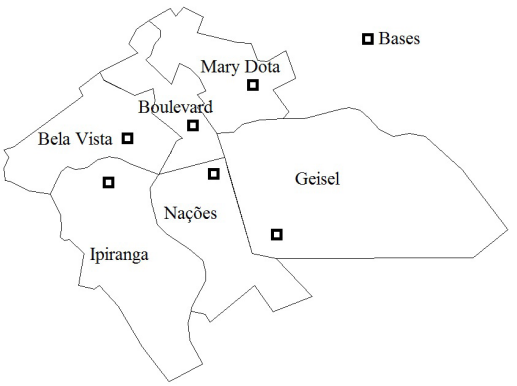

Figura 5. Mapa da cidade de Bauru com seus átomos e bases.

O SAMU Bauru dividia os chamados de acordo com o tipo de urgência do chamado recebido: vermelho, chamados urgentes em que há risco de vida para o paciente; amarelo, emergência grave; verde, emergência moderada; azul, emergência leve. Para modelagem em *layering*, cada átomo foi dividido em subátomos (*layers*).

4.2 Processo de chegada

Foi analisado se o intervalo entre as chegadas se aproximava de uma distribuição exponencial, o teste de aderência realizado foi o de Kolmogorov Smirnov, o resultado está representado na Figura 6. Considerando-se um nível de significância de 5%, a hipótese de que os dados seguiam uma distribuição exponencial não foi rejeitada, observando-se um valor-p de 0,172.

O cálculo da taxa de chegada individual de cada subátomo é feito pelo produto da proporção dos chamados da subárea (p_{jk}) e a taxa de chegada do sistema. A Tabela 3 mostra a identificação de cada subátomo, de acordo com a região e sua prioridade, e as respectivas taxas de chegada de usuários de 2013.

4.3 Tempo de viagem

Os tempos médios de viagem entre os átomos foram estimados a partir da amostra. Nos casos em que não houve observação no período analisado, utilizou-se a ferramenta Google Earth® para se estimar esses dados. Com a ferramenta foi possível calcular as distâncias entre os centroides dos átomos: assim, supondo-se uma velocidade de trânsito média de 60 km/h, estimou-se o tempo médio de viagem (Tabela 4). Os dados indicados por asterisco (*) são os obtidos por meio dessas estimativas.

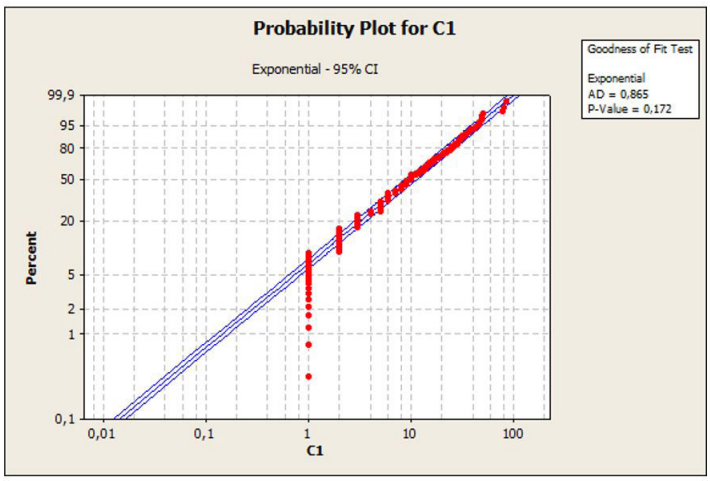

Figura 6. Teste de aderência para a distribuição exponencial. Fonte: Autores.

Tabela 3. Identificação dos átomos e subátomos e suas taxas de chegada.

Subárea	Átomo	Subátomo	Nº de chamados	p_{jk}	λ_{jk} (chamados/hora)
Geisel azul	1	1*d*	1	0,0050	0,0174
Geisel verde	1	1*c*	12	0,0594	0,2090
Geisel amarelo	1	1*b*	15	0,0743	0,2612
Geisel vermelho	1	1*a*	5	0,0248	0,0871
Nações azul	2	2*d*	5	0,0248	0,0871
Nações verde	2	2*c*	20	0,0990	0,3483
Nações amarelo	2	2*b*	23	0,1139	0,4005
Nações vermelho	2	2*a*	3	0,0149	0,0522
Ipiranga azul	3	3*d*	3	0,0149	0,0522
Ipiranga verde	3	3*c*	11	0,0545	0,1916
Ipiranga amarelo	3	3*b*	9	0,0446	0,1567
Ipiranga vermelho	3	3*a*	4	0,0198	0,0697
Mary Dota azul	4	4*d*	2	0,0099	0,0348
Mary Dota verde	4	4*c*	15	0,0743	0,2612
Mary Dota amarelo	4	4*b*	7	0,0347	0,1219
Mary Dota vermelho	4	4*a*	2	0,0099	0,0348
Bela Vista azul	5	5*d*	7	0,0347	0,1219
Bela Vista verde	5	5*c*	18	0,0891	0,3134
Bela Vista amarelo	5	5*b*	23	0,1139	0,4005
Bela Vista vermelho	5	5*a*	6	0,0297	0,1045
Boulevard azul	6	6*d*	1	0,0050	0,0174
Boulevard verde	6	6*c*	3	0,0149	0,0522
Boulevard amarelo	6	6*b*	6	0,0297	0,1045
Boulevard vermelho	6	6*a*	1	0,0050	0,0174
SAMU Bauru			202	1,0000	3,5176

4.4 Servidores

Conforme mencionado anteriormente, o sistema do SAMU Bauru é composto por 9 ambulâncias na cidade de Bauru. Sendo que, dessas, 2 são Veículos de Serviço Avançado (VSA) e 7 são Veículos de Serviço Básico (VSB). Além disso, todas as ambulâncias podem transitar entre os átomos, o que confirma a hipótese do modelo hipercubo.

4.5 Localização dos servidores

Cada servidor está localizado em uma base fixa. Os VSA ficam na sede do SAMU Bauru, na região Geisel, enquanto os VSB se encontram espacialmente distribuídas da seguinte maneira: 1 no Geisel, 1 na Nações, 1 na Ipiranga, 1 no Mary Dota, 2 na Bela Vista e 1 na Boulevard.

4.6 Despacho dos servidores

Foi possível notar que, na grande maioria dos casos, cerca de 96%, os chamados foram atendidos por apenas uma ambulância. Apenas em alguns casos, menos de 4%, houve o envio de mais de um veículo para a cena da ocorrência. Além disso, os VSA atendem apenas chamados vermelhos (graves), isso não está de acordo com a hipótese para o modelo hipercubo utilizado, mas a carga de trabalho sobre o sistema é baixa, permitindo uma flexibilização dessa hipótese, como visto em Takeda et al. (2004, 2007). A formação de fila de espera é permitida e não possui restrições quanto ao tamanho, sendo que os chamados em fila são organizados de acordo com sua prioridade.

4.7 Preferência de despacho

A política de despacho dos servidores leva em conta alguns aspectos: localização do chamado, tipo de urgência e localização dos servidores. A preferência é dada ao servidor da região da ocorrência, caso esse esteja ocupado é enviado o servidor mais próximo, todavia para chamados vermelhos é enviado um dos VSA, escolhido aleatoriamente. Esses aspectos causam aleatoriedade para o despacho, ou seja, não é possível escrever uma matriz de preferência de despacho (Burwell et al., 1993; Takeda et al., 2007).

4.8 Tempo de serviço

Os tempos de serviço de cada servidor foram analisados a fim de se verificar se eram exponencialmente distribuídos. Para tanto, utilizou-se o método Kolmorogov-Smirnov – para todos os servidores, a hipótese foi rejeitada ao nível de significância de

Tabela 4. Tempos médios de viagem entre subátomos (minutos). Os dados indicados por asterisco (*) são os obtidos por meio de estimativas considerando as distâncias dos centroides dos átomos e supondo-se uma velocidade de trânsito média de 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*

5%. Além disso, foi verificada a homogeneidade dos servidores, para o que se realizou a Análise de Variância (ANOVA) ao nível de significância de 5%. Os resultados mostraram diferenças entre as médias dos tempos de serviço de todos os VSB, exceto os localizados na Bela Vista, de forma que os servidores são heterogêneos, exceto os VSA e os VSB da Bela Vista.

A Tabela 5 mostra as taxas de serviço de cada servidor, que foram calculadas a partir dos seus tempos médios de serviço.

4.9 Relação entre o tempo de serviço e o tempo de viagem

A Tabela 6 faz a comparação entre o tempo de serviço e o tempo médio de viagem. Nota-se que os tempos médios de viagem são relativamente pequenos com relação aos tempos médios de atendimento, sendo que a relação, em nenhum dos casos, excedeu 25%, validando essa hipótese para utilização do modelo hipercubo.

5 Aplicação do modelo e análise dos resultados do SAMU Bauru

O modelo utilizado para a aplicação de teoria das filas foi o modelo hipercubo, visto que o sistema conseguiu atender à maioria das hipóteses do modelo clássico e a distinção dos servidores é necessária. Devido à não aderência à todas as hipóteses, foi feito o uso de extensões para o modelo hipercubo clássico. Foram escolhidas a aleatoriedade no despacho e a prioridade em fila, conforme dito na seção 3, de acordo com as características do SAMU Bauru encontradas durante a verificação das hipóteses, durante a qual os chamados foram organizados de acordo com a gravidade e o envio das ambulâncias não obedeceu uma lista de preferência de despacho fixa.

Vale ressaltar que a técnica para inclusão da aleatoriedade no despacho utilizada foi a de criação de matrizes de preferências de despacho de forma aleatória. Para a construção do modelo utilizou-se a linguagem de programação Pascal. Para a resolução do modelo foi realizado um total de 50 replicações e, quando necessário, foi realizada a calibração dos tempos de serviço considerando-se tolerância de 0,1 chamados/hora.

Para o modelo inicial (original), utilizado para validação do modelo, foram comparadas as cargas de trabalho dos servidores e os tempos de viagem. Não foram comparados os tempos de espera em fila, pois não foi possível obter dados referentes aos tempos de triagem dos chamados, o que influencia o tempo de espera da amostra.

5.1 Cenário original *versus* amostra

A partir da resolução do sistema foi possível realizar a comparação dos resultados obtidos com as informações retiradas da Amostra.

Tabela 5. Taxas médias de atendimento para cada ambulância.

Ambulâncias		Tempo médio de serviço (min.)	μ_i (chamados/hora)
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

Tabela 6. Relação entre o tempo médio de atendimento e o tempo médio de viagem para os servidores.

Ambulâncias		Tempo médio de serviço (min.)	Tempo médio de viagem (min.)	Relação
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

A Tabela 7 mostra a carga de trabalho (*workload*) dos servidores no sistema modelado em comparação com os resultados da amostra e seus desvios relativos. Essa medida foi considerada como tendo boa aderência à amostra, com um desvio médio em torno de 8%. Pode-se destacar que a carga de trabalho das ambulâncias avançadas (GA1 e GA2) teve um acréscimo (7,4%) em relação à amostra, o que pode ser explicado por o modelo não restringir os atendimentos dessas ambulâncias aos chamados mais graves. Contudo, esse desvio não compromete a validação do modelo nessa comparação. A ambulância do átomo Mary Dota (MD) teve o maior erro, 15%, o qual foi considerado aceitável, levando-se em conta a média dos resultados.

Os tempos médios de viagem para cada subátomo são mostrados na Tabela 8. O desvio médio encontrado foi de 9%. Por outro lado, devido ao pequeno tamanho da amostra para cada átomo (5 amostras ou menos para 11 dos 24 subátomos), os dados dos átomos *b*, *c* e *d* foram agregados, diminuindo o desvio encontrado, sem perda de precisão, visto que são atendidos pelas mesmas ambulâncias e possuem mesma preferência de despacho. Os maiores desvios, mas ainda considerados aceitáveis, referiram-se aos subátomos de prioridade *a*, que possuem uma amostra menor e uma menor taxa de chegada no sistema (aproximadamente 10% dos chamados do sistema) e não possibilitam agregar dados com outras prioridades.

Por último, foi feita a análise dos tempos médios de viagem dos servidores. A Tabela 9, mostra a comparação desses tempos em relação à amostra e ao modelo. Nesse caso observa-se uma melhor aderência do modelo, visto que o desvio médio foi inferior à 6%. O tamanho da amostra favorece a obtenção desse resultado, já que todos os servidores tiveram pelo menos 20 chamados na amostra, alguns inclusive ultrapassando 50 (BV1 e BV2). O maior desvio observado (ambulância NÇ) foi no caso com menor número de amostras, com 21 chamados.

Um outro dado que mostra a aderência do modelo à amostra é o tempo médio de viagem do sistema, o qual teve um valor de 9,7 minutos, com um desvio de 1% em relação à amostra. Além disso, não foi necessário realizar a calibragem do tempo médio de serviço, o que também mostra uma boa aderência do modelo.

Como os resultados das cargas de trabalho e dos tempos médios de viagem mostram uma boa aderência do modelo ao sistema SAMU Bauru, considerou-se o modelo validado, possibilitando a análise dos cenários alternativos. Inicialmente foram estudados os efeitos do aumento na demanda do sistema baseando-se em prospecções dos dados do SAMU Bauru de 2012 e 2013 e aumentos atípicos. Depois avaliou-se a inclusão de uma nova ambulância para mitigar os efeitos no aumento na demanda prevista no médio e longo prazos.

5.2 Aumento na demanda

Para o aumento na demanda, encontrou-se a tendência do sistema para 2014 (ano seguinte aos dos dados coletados). Para tanto, foram utilizados três meios de cálculo para a previsão:

- Proporção dos chamados (13,57%): Encontrou-se o aumento da proporção dos chamados no período de pico de 2012 e 2013 para o sistema de Bauru como um todo, e supôs-se que essa taxa de aumento se manteria para 2014;
- Análise de série temporal (5,71%): Através do resumo dos chamados mês a mês de 2012 e 2013, buscou-se encontrar previsões para o ano de 2014, até o período de pico, por meio do método da decomposição, em que foi escolhido o mês de setembro, arbitrariamente, para estudo do aumento na demanda;
- Outras previsões (25% e 50%): Escolha arbitrária de aumentos buscando entender situações atípicas extremas.

Os cenários de aumento na demanda mostram o impacto sobre o sistema causado pelo aumento no número médio de usuários que chegam ao sistema, de forma a congestioná-lo. Isso aumenta os tempos de espera em fila e o número de atendimentos realizados

Tabela 7. Comparação entre as cargas de trabalho obtidas pelo modelo e pela amostra.

Ambulância	Amostra	Modelo	Desvio relativo
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%

Tabela 8. Comparação entre os tempos médios de viagem (min.) aos subátomos obtidos pelo modelo e pela amostra.

Subátomos	Amostra	Modelo	Desvio relativo
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%

Tabela 9. Comparação entre os tempos médios de viagem (min.) dos servidores obtidos pelo modelo e pela amostra.

Ambulância	Amostra	Modelo	Desvio relativo
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%

pelos *backups* (ambulâncias enviadas caso a prioritária esteja ocupada).

A Figura 7 mostra esse impacto sobre os tempos médios de resposta das ambulâncias. É importante observar 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.

O impacto também é observado sobre os usuários do sistema. A Figura 8 mostra os diferentes efeitos sobre cada uma das classes de usuários do sistema. Os usuários de maior prioridade possuem o maior tempo de resposta nos cenários iniciais, já que são atendidos pelos VSA e esses muitas vezes precisam cruzar o sistema para atendê-los. Porém eles sofrem menos com os aumentos na demanda. Os usuários de menor prioridade, atendidos pelos VSB locais, possuem um baixo tempo de resposta, cerca de 11 minutos nos cenários iniciais. Contudo, eles sofrem mais com os aumentos na demanda, chegando a uma média de 34 minutos para os chamados verdes e azuis. Isso mostra também a importância de se representar as diferentes prioridades de chamados.

5.3 Inclusão de nova ambulância

Para cada um dos cenários de aumento na demanda foram analisados o impacto, nas medidas de desempenho, de se incluir uma nova ambulância (VSB) no sistema. Essa análise é importante pois

podemos medir o quanto e aonde a inclusão de um servidor pode melhorar o desempenho do sistema. A adição de ambulância foi testada em cada uma das bases (átomos) já existentes do SAMU Bauru. A nova ambulância utilizou os mesmos dados (tempos de atendimento e localização) que as ambulâncias já localizadas no átomo em que foi inserida. A avaliação da mitigação dos efeitos do aumento na demanda foi feita através do tempo médio de resposta do sistema.

A Figura 9 compara o impacto da inclusão de uma nova ambulância no tempo médio de resposta. Note-se que o modelo consegue mostrar a melhor localização para a nova ambulância e também quantificar o efeito sobre as medidas de desempenho do sistema. Dessa forma, caso o gestor do sistema queira estar preparado para um período com aumento de 25% na demanda, pode manter o mesmo tempo médio de resposta com a adição de apenas uma nova ambulância no átomo Boulevard, mitigando os efeitos do aumento na demanda.

A nova ambulância trouxe melhores resultados em todos os cenários quando presente no átomo Boulevard,

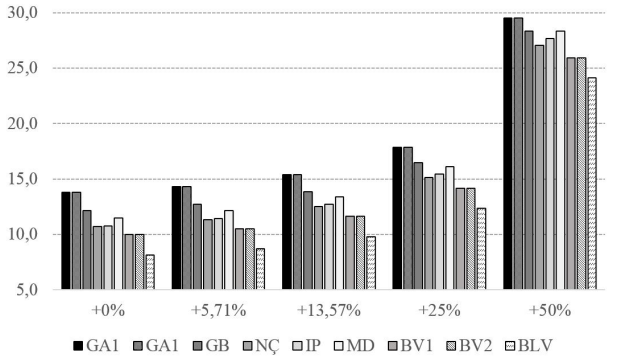

Figura 7. Impacto do aumento na demanda sobre os tempos de resposta das ambulâncias (min.).

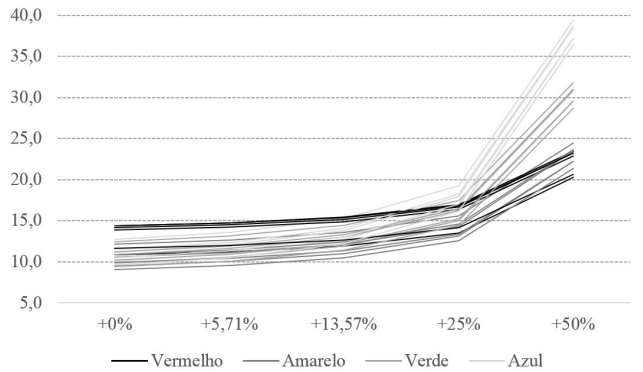

Figura 8. Impacto do aumento na demanda sobre os tempos de resposta aos subátomos (min.).

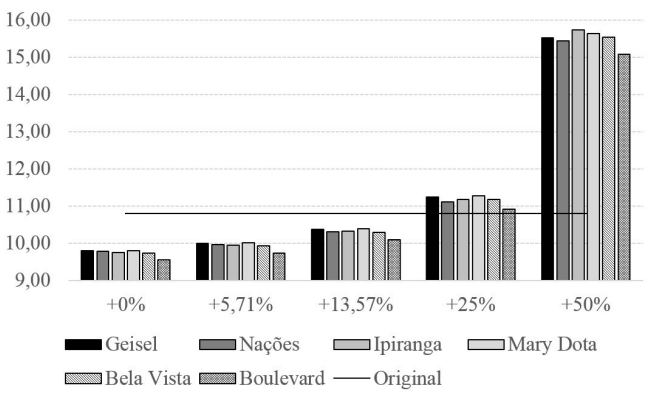

Figura 9. Comparação dos tempos médios de resposta com a inclusão de nova ambulância (min.).

obtendo-se um tempo médio de resposta 3% inferior às demais localidades, em média. Por outro lado, colocá-la no átomo Mary Dota trouxe os resultados menos vantajosos nos quatro primeiros cenários, com uma redução de 9% contra 12% no átomo Boulevard, no cenário original. No cenário mais crítico, com aumento de 50%, as duas melhores opções para a localização da ambulância são os átomos Nações e Boulevard, ambos em regiões centrais da cidade.

6 Conclusões

O SAMU Bauru é um sistema que atende a população em quatro categorias de prioridades. Aa qualidade do serviço está intimamente ligada a uma boa gestão dos recursos disponíveis. Dessa forma, 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 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.

Consideramos os desvios dos dados da amostra, comparados com o modelo, aceitáveis, possibilitando a análise dos cenários alternativos. Os cenários alternativos avaliaram o impacto do aumento de demanda no sistema e a melhor opção para inclusão de uma nova ambulância levando em conta aumentos na demanda. O modelo fornece uma visão analítica e detalhada do impacto das decisões tomadas no gerenciamento de um SAE de forma a possibilitar melhor atendimento ao público com melhor tempo de resposta e envio de ambulâncias apropriadas. Faz isso porque permite avaliação e prospecção de situações futuras quando devidamente calibrado ao sistema em análise.

Na análise dos cenários foi possível observar 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%, para os quais o acréscimo nos tempos de resposta foi 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 deram quando ela estava presente no átomo Boulevard, obtendo-se então um tempo médio de resposta 3% inferior a quando locada nas demais localidades, em média. Por outro lado, colocá-la no átomo Mary Dota trouxe os resultados menos vantajosos nos quatro primeiros cenários, com uma redução de 9% contra 12% no átomo Boulevard no cenário original. No cenário mais crítico, aumento de 50%, as duas melhores opções para a localização dessa nova ambulância são os átomos Nações e Boulevard, ambos em regiões centrais da cidade.

Para pesquisas futuras propõe-se a utilização de modelo com *backup* parcial, com filas e restrições aos tipos dos chamados, como constatou-se na avaliação dos VSA na amostra, na qual eles atendem apenas os chamados mais graves, em qualquer localização do sistema. Além disso, propõe-se que se avalie a localização dinâmica dos servidores em função do tempo, a partir de um modelo dinâmico, de forma a captar o efeito das variações que o sistema sofre ao longo do dia.

Agradecimentos

Os autores agradecem à CAPES e à FAPESP pelo financiamento da pesquisa. Também agradecem ao SAMU de Bauru pela disponibilização dos dados e ao revisor anônimo pelos preciosos comentários, que foram muito valiosos para a evolução do trabalho.

Referências

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.