Production, v. 24, n. 1, p. 1-12, jan./mar. 2014
doi: http://dx.doi.org/10.1590/S0103-65132013005000028

# Extensão do modelo hipercubo para análise de sistemas de atendimento médico emergencial com prioridade na fila

**Regiane Máximo de Souza[a*], Reinaldo Morabito[b], Fernando Y. Chiyoshi[c], Ana Paula Iannoni[d]**

[a*]regiane@feb.unesp.br, UNESP, Brasil
[b]morabito@ufscar.br, UFSCar, Brasil
[c]chiyoshi@pep.ufrj.br, UFRJ, Brasil
[d]iannoni93@hotmail.com, Ecole Centrale Paris, França

## Resumo

Em alguns sistemas de atendimento médico emergencial, a demanda pelo serviço pode ser alta devido ao atendimento a pacientes em diferentes estados, desde mais graves até mais leves. Nesses sistemas, pode haver formação de filas de usuários aguardando atendimento, e a necessidade de se considerar explicitamente políticas de prioridade nesse atendimento torna-se importante. Neste trabalho propõe-se uma extensão do clássico modelo hipercubo de filas espacialmente distribuídas para considerar fila com prioridade. Para verificar a viabilidade e a aplicabilidade dessa abordagem, utilizam-se dados de um estudo de caso realizado no SAMU de Ribeirão Preto-SP. Foram analisados dois cenários que consideram dois aspectos relevantes: o impacto dos atendimentos de remoção de pacientes e o aumento da demanda nas diversas classes de chamados dos usuários do sistema. O foco é no tempo médio de resposta aos chamados dos usuários, considerado como uma medida de desempenho importante do sistema, principalmente aos chamados de classes com alta prioridade. Os resultados mostram que a abordagem pode ser utilizada para analisar satisfatoriamente sistemas com prioridade de fila.

## Palavras-chave

Modelo hipercubo. Prioridade na fila. SAMU. Atendimento médico emergencial.

## 1. Introdução

Em sistemas de atendimento médico emergencial como os SAMUs (Sistema de Atendimento Móvel de Urgência) no Brasil, o tempo de resposta aos usuários é de fundamental importância, pois, dependendo do estado do paciente, a demora no atendimento pode significar sua vida ou morte. Devido às restrições orçamentárias, em geral os SAMUs não dispõem de um grande efetivo de pessoas e equipamentos, incluindo ambulâncias e suas tripulações, e, portanto, os gestores desses sistemas precisam tomar decisões de planejamento e operação destes, levando em conta diferentes conflitos (*trade-offs*) entre investimentos, custos operacionais e níveis de serviço oferecidos aos usuários. No Brasil, o SAMU é um programa do governo federal que tem a finalidade de prestar socorro médico emergencial às pessoas e garantir a qualidade no atendimento. Esse serviço presta atendimento em qualquer local, como residências, locais de trabalho e vias públicas. Esses sistemas possuem particularidades que os distinguem entre si, como, por exemplo, diferenças em relação aos tipos de chamados dos usuários, tipos, números e localizações das ambulâncias, regras de despacho das ambulâncias etc., que devem ser cuidadosamente estudadas e consideradas na análise e configuração desses sistemas.

O modelo hipercubo de filas espacialmente distribuídas, proposto originalmente por Larson (1974) e estendido por vários autores (SWERSEY, 1994; GALVÃO; MORABITO, 2008; BOFFEY; GALVÃO; ESPEJO, 2007), tem sido utilizado para analisar e planejar diversos sistemas de atendimento emergencial.

*UFSCar, São Carlos, SP, Brasil
Recebido 16/03/2011; Aceito 27/09/2012

O modelo original considera disciplinas de prioridade no despacho dos servidores, se houver servidores disponíveis para despacho imediato, e fila com disciplina de atendimento FCFS (*First Come First Served*), no caso de todos os servidores estarem ocupados. Alguns exemplos de extensões e/ou aplicações do modelo hipercubo em sistemas de atendimento emergencial aparecem em Atkinson et al. (2006, 2008), Chelst e Barlach (1981), Brandeau e Larson (1986), Burwell, Jarvis e McKnew (1993), Sacks e Grieff (1984), Swersey (1994), Marianov e Serra (1998, 2003), Marianov e Ríos (2000), Mendonça e Morabito (2001), Chiyoshi, Galvão e Morabito (2000), Larson e Odoni (2007), Takeda, Widmer e Morabito (2007), Iannoni e Morabito (2007), Iannoni, Morabito e Saydam (2009) e Morabito, Chiyoshi e Galvão (2008).

Em sistemas de prestação de serviços, a necessidade de utilizar prioridades pode ocorrer em relação a servidores ou usuários. Em modelos hipercubos de filas, prioridades relativas a servidores podem ser tratadas através de recurso conhecido como *layering* (LARSON; ODONI, 2007), permitindo assegurar que, preferencialmente, unidades com recursos adequados sejam despachadas de acordo com as necessidades dos usuários. No presente estudo, consideramos prioridades associadas a usuários em que se permite o atendimento de usuários prioritários antes de usuários comuns, independentemente da ordem de chegada.

Na abordagem de *layering*, as classes de chamados de usuários são diferenciadas e as diferentes prioridades são representadas no modelo hipercubo subdividindo os átomos de demanda em camadas (*layers*). Essa abordagem foi utilizada, por exemplo, em Takeda, Widmer e Morabito (2007) ao estudar o SAMU da cidade de Campinas-SP, subdividindo cada átomo geográfico do sistema em dois subátomos (não geográficos), para representar duas classes de chamados: usuários mais graves, que correm risco de vida, e usuários menos graves, que não correm risco de vida. O estudo mostrou que, mesmo sem considerar explicitamente prioridade na fila (com disciplina de atendimento FCFS, conforme mencionado acima), mas apenas prioridade no despacho das ambulâncias quando uma ou mais estão disponíveis para atendimento, o modelo hipercubo foi adequado para analisar sistemas como o SAMU-Campinas.

O SAMU de Ribeirão Preto (SAMU-RP) atende a chamados de grave a leve e realiza atendimentos de remoção de pacientes, caracterizado pelo transporte de pacientes, na maioria agendados. Ao desconsiderar os atendimentos de remoção, Souza et al. (2010) mostraram que o modelo hipercubo clássico (com a mesma abordagem de *layering* utilizada em Takeda, Widmer e Morabito (2007) foi adequado para analisar também o SAMU-RP, principalmente em situações em que a taxa de utilização do sistema é bem menor de 50%. No entanto, em situações de maior congestionamento do sistema, em função de possíveis aumentos nas demandas dos usuários, ou quando o agendamento da remoção de pacientes compromete a disponibilidade das ambulâncias, o estudo mostrou a importância de estender o modelo hipercubo clássico para analisar sistemas com diferentes classes de usuários, incorporando políticas específicas de prioridade na fila para atendê-los. Para isso, resultados do modelo hipercubo clássico foram comparados com os de um modelo de simulação discreta considerando prioridade de fila para o mesmo sistema.

Nesse contexto, o objetivo do presente trabalho é estender o modelo hipercubo clássico para tratar explicitamente políticas específicas de prioridade na fila. À medida que o nível de utilização do sistema cresce, a probabilidade de fila de usuários torna-se bem maior que zero, e o tempo médio de resposta aos usuários também cresce em função do aumento no tempo médio de espera na fila, além do tempo médio de viagem. Essa extensão do modelo hipercubo é especialmente importante para os usuários das classes de maior prioridade, que apresentam maior risco de vida.

A literatura relativa a modelos hipercubos de filas é predominantemente composta de trabalhos que se originaram no estudo de sistemas reais de prestação de serviços de emergência, em geral envolvendo serviço médico, combate ao incêndio e combate ao crime. Esses sistemas comumente operam em regime de 24 horas por dia e 7 dias por semana, com taxas de ocupação relativamente baixas devido à importância do tempo de resposta como medida da qualidade de serviço. Pode-se dizer que o estado normal desses sistemas é o estado estacionário, o estado de equilíbrio (*steady state*). O problema que originou o nosso trabalho não se afasta do padrão encontrado na literatura envolvendo serviço médico de emergência. A metodologia de análise adotada atende às necessidades do problema estudado, concentrando-se na determinação das características operacionais do sistema no estado estacionário.

Este trabalho está organizado da seguinte maneira: a seção 2 descreve a extensão do modelo hipercubo para considerar prioridade na fila, assim como as modificações para cálculo das medidas de desempenho. A seção 3 apresenta um exemplo ilustrativo resolvido com a abordagem proposta. A seção 4 mostra a aplicação do modelo hipercubo estendido para análise de dois cenários baseados em dados do SAMU-RP; o primeiro avalia o impacto do atendimento de remoção de pacientes e o segundo avalia o impacto do aumento de demanda das classes de usuários no sistema. Finalmente, a seção 5 apresenta as conclusões do estudo e discute algumas perspectivas para pesquisa futura.

## 2. Prioridades em modelo hipercubo de filas

O modelo hipercubo é um modelo descritivo utilizado como ferramenta para análise e planejamento de sistemas de emergência urbanos. Além de considerar incertezas quanto à origem dos chamados, tempos de serviço e disponibilidade dos servidores, o modelo aborda complexidades geográficas e temporais da região, com base em filas espacialmente distribuídas. Ele pode analisar tanto sistemas coordenados como centralizados – quando o usuário liga para uma central solicitando algum tipo de serviço e um servidor se desloca até o cliente. Basicamente, a ideia é expandir o espaço de estados de um sistema de fila $M/M/N$ (em que $N$ é o número de servidores) a fim de representar cada servidor individualmente, podendo incluir políticas de despacho mais complicadas. A solução do modelo é dada partindo-se da construção do conjunto de equações de equilíbrio para o sistema. Os resultados baseiam-se nos valores das probabilidades de equilíbrio de estado do sistema, possibilitando o cálculo de 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. Algumas dessas hipóteses podem ser alteradas, como, por exemplo, múltiplo despacho e *backup* parcial, como em Chelst e Barlach (1981), Mendonça e Morabito (2001), e Iannoni et. al. (2006, 2007, 2009).

Quando se considera o modelo hipercubo em sua forma básica, os estados do sistema pertencem ao hipercubo (formado pelos $2^N$ vértices do sistema) ou à fila, com exceção do estado $\{11... 1\}$, em que todos os servidores estão ocupados e não há fila de espera, que pertence a ambos os subsistemas. É através desse estado que se estabelece a comunicação entre os estados do hipercubo e da fila. Para considerar prioridades no atendimento aos clientes em espera, os estados da fila devem ser definidos de forma a detalhar sua constituição, não sendo suficiente utilizar apenas o tamanho da fila (i. e., número de usuários) para definir os estados da fila, como na formulação básica em sistemas sem prioridades. Note-se que a forma expandida de definir os estados do sistema é requerida apenas para os estados da fila, de modo que o tratamento de prioridades no atendimento pode ser analisado a partir do estado $\{11... 1\}$, em que o sistema está ocupado mas a fila está vazia.

Para um sistema com $r$ classes de prioridades, os estados da fila com $n$ usuários podem ser representados por cadeias ordenadas de $n$ elementos, que indicam as classes de prioridade dos usuários. Assim, os estados de uma fila com $n = 2$ usuários em um sistema com $r = 3$ classes de prioridade $(a,b,c)$ serão: $\{aa\}$, $\{ab\}$, $\{ac\}$, $\{bb\}$, $\{bc\}$ e $\{cc\}$. A Figura 1 mostra os estados da fila com até três usuários para um sistema com $r = 3$ classes de prioridade e $N = 3$ servidores, bem como as possíveis transições entre os estados. A notação $S_m$ é o conjunto de estados da fila com m usuários no sistema (i. e., $n = m - N$ usuários na fila).

De forma similar aos estados dos vértices do hipercubo (estados sem fila), as equações de equilíbrio para os estados na fila são construídas igualando-se a taxa média de transição do sistema para cada estado e a taxa média de transição do sistema para fora do mesmo estado. A Figura 2 mostra as taxas de transição em torno do estado $\{ab\}$ onde $\lambda_k$ , $k \in D = \{a,b,c\}$

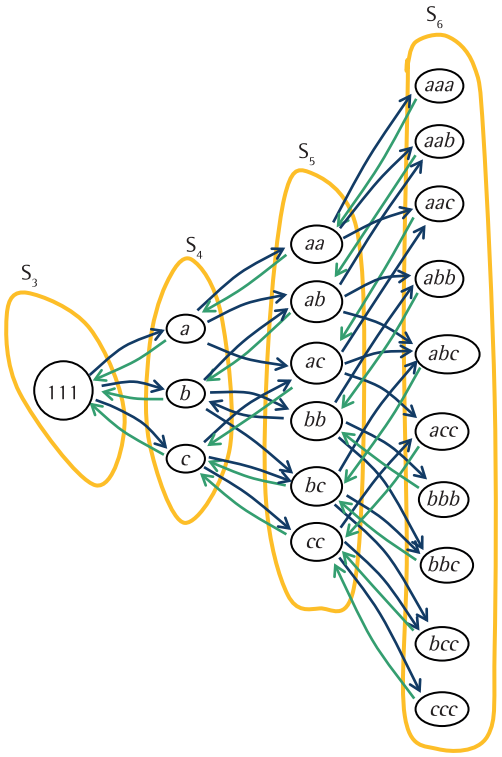


**Figura 1.** Estados da fila em um sistema com até três usuários em fila e três classes de prioridades e três servidores.

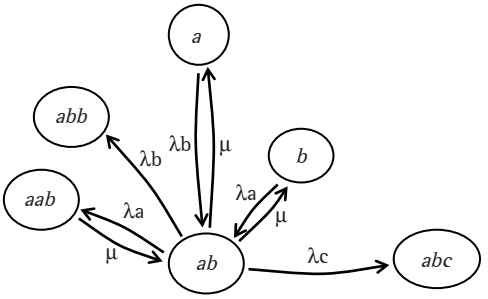


**Figura 2.** Vértice $\{ab\}$ e seus adjacentes.

representa a taxa de chegadas dos usuários da classe de prioridade $k$ e $\mu$ a taxa de atendimento do sistema. A equação de equilíbrio correspondente é dada por:

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

em que $\lambda = \sum_{k \in D} \lambda_k$, e $P_{\{ab\}}$ é a probabilidade de equilíbrio do estado $\{ab\}$. As equações de equilíbrio relativas aos demais estados da fila podem ser obtidas utilizando-se o mesmo procedimento.

As equações de equilíbrio relativas aos estados da fila com um usuário, por exemplo, para o estado $\{a\}$:

$$(\lambda+\mu)P_{\{a\}} = \mu P_{\{aa\}} + \lambda_a P_{\{111\}}$$

merecem especial atenção. Note-se que elas contêm o termo $P_{\{111\}}$ que, embora represente a probabilidade de um estado da fila, não é considerada uma variável neste contexto, de modo que as equações de equilíbrio formam um sistema não homogêneo de $K$ equações a $K$ incógnitas, sendo $K$ o número de estados da fila a menos do estado $\{111\}$. De forma geral, pode-se encontrar o número de estados da fila para $r$ classes de prioridade e $n$ usuários em espera por:

$$f(n,r) = \binom{r+n-1}{n}$$

de modo que o número total de estados ($K$) com filas para $r$ classes e limite de fila $L$ é:

$$K = \sum_{n=1}^{L} f(n,r) = \binom{r+L}{L} - 1.$$

A probabilidade $P_{\{111\}}$ é determinada a partir das $2^N$ equações relativas aos $2^N$ estados (vértices) do hipercubo (LARSON; ODONI, 2007; CHIYOSHI; GALVÃO; MORABITO, 2000). Inicialmente resolvemos as equações do hipercubo supondo espaço zero de espera. A seguir adicionamos um estado $Q$ representando a fila de espera, utilizando a relação de equilíbrio $\lambda P_{\{11...1\}} = \mu P\{Q\}$ entre os estados $\{11...1\}$ e $\{Q\}$. O conjunto de números resultante não constitui uma distribuição de probabilidade, pois $P_{\{00..0\}} + ... + P_{\{11..1\}} = 1 - P_{\{Q\}}$. Quando esse conjunto é normalizado para obter-se uma distribuição de probabilidade, os valores originais das probabilidades dos estados do hipercubo são alterados. Esse fato suscita dúvida quanto à validade dos novos números como soluções do sistema original de $2^N$ equações de equilíbrio, relativas aos estados do hipercubo sem espaço de espera. Para dirimir essa dúvida, lembramos que as equações de equilíbrio dos estados do hipercubo estabelecem as relações entre as probabilidades e constituem um sistema de equação homogêneo (indefinido). A equação adicional que utilizamos para eliminar a indefinição é a equação de normalização das probabilidades para soma um. Esse procedimento conduz a uma solução particular do sistema de equações do hipercubo que constitui uma distribuição de probabilidade. Quando incluímos o estado $\{Q\}$, a normalização do conjunto aumentado equivale à normalização das probabilidades originais para soma $1 - P_{\{Q\}}$, de modo que os valores modificados são também soluções do sistema original de equações.

Pela solução do sistema de equações de equilíbrio, obtêm-se as probabilidades associadas aos estados da fila, que permitem determinar as características operacionais do sistema. Em particular, o tamanho médio da fila para usuários da classe de prioridade $k$ pode ser obtido através de:

$$L_{qk} = \sum_{n_k=1}^{a} n_k \sum_{S \in E_{n_k}^{[k]}} P(S)$$

onde $E_{n_k}^{[k]}$ é o conjunto de estados com $n_k$ usuários da classe $k$ e $\alpha$ é o tamanho máximo da fila. Os tempos médios de espera podem ser obtidos utilizando-se a fórmula de Little (1961). Essa fórmula define a relação entre o tamanho médio da fila e o tempo médio de espera na fila. Apesar de sua importância, a literatura básica sobre teoria das filas é econômica em relação à fórmula de Little. Em face da relevância de sua função na metodologia utilizada no presente trabalho, julgamos oportuno apresentar uma breve discussão do tema. A derivação da fórmula de Little a seguir baseia-se na derivação dada por Whitt (1991).

Definindo $A_i$ como o instante de entrada do usuário $i$ na fila e $D_i$ como o instante de saída do usuário $i$ da fila, podemos representar o tempo de espera $W_i = (D_i - A_i)$ do usuário $i$ no plano usuário-tempo através de uma barra horizontal de altura unitária localizada entre $(i, A_i)$ e $(i, D_i)$. Para um intervalo de tempo $[0, T]$ em que $n$ usuários entram e saem da fila, o tempo total de espera dos usuários é dado pela área total das barras $\sum_{i=1}^{n} W_i$. Em relação à fila, se definirmos $Q(t)$ como o tamanho da fila no instante $t$, vê-se que $Q(t)$ é dado pelo número de barras interceptadas por uma linha vertical em $t$. A fila total relativa aos $n$ usuários é dada por $\int_{t=0}^{T} Q(t)\,dt$, sendo representada pela mesma área que representa a espera total dos usuários. Em relação a médias em $[0, T]$, temos $L^{(T)} = (1/T)\int_{t=0}^{T} Q(t)\,dt$ para o tamanho da fila e $W^{(n)} = (1/n)\sum_{i=1}^{n} Wi$ para o tempo de espera. A equivalência:

$$\sum_{i=1}^{n} W_i = \int_{t=0}^{T} Q(t)\,dt$$

entre espera total e fila total permite escrever:

$$L^{(T)}T = nW^{(n)} \; ou \; L^{(T)} = (n/T)W^{(n)} \; ou \; L^{(T)} = \lambda^{(n)}W^{(n)}$$

onde $\lambda^{(n)} = (n / T)$ é a taxa média de chegada de usuários em $[0,T]$. Essa é a derivação da fórmula de Little para uma situação particular em que se consideram $n$ usuários que entram e saem da fila em $[0,T]$. Para uma formulação mais geral, deve-se considerar o sistema em um intervalo de tempo genérico $[0,t]$, em que podem existir usuários que entram na fila antes de t e saem da fila depois de $t$. Para um usuário particular $j$ nessa situação, tem-se $A_j < t < D_j$. Como o total da fila é dada pela área definida pela função $Q(t)$ em $[0,t]$, a integral $\int_{t=0}^{t} Q(s)ds$ inclui apenas a parcela do tempo de espera do usuário $j$ de $A_j$ até $t$. Por outro lado, o tempo total de espera resulta da soma das esperas individuais dos usuários, não comportando a inclusão parcial do tempo de espera de qualquer usuário. A fila total $\int_{t=0}^{t} Q(s)ds$ pode ser aproximada inferiormente por $\sum_{i=1}^{D(t)} W_i$ , sendo $D(t)$ o número total de usuários que saíram da fila até $t$, excluindo o tempo de espera do usuário $j$, e aproximada superiormente por $\sum_{i=1}^{A(t)} W_i$ , sendo $A(t)$ o número total de usuários que entraram na fila até $t$, incluindo o tempo de espera do usuário $j$. Note-se que o caso particular descrito acima para o atendimento de $n$ usuários em $[0,T]$ corresponde à situação em que $A(t) = D(t)$ (o sistema está vazio em $T$). De uma forma geral podemos então escrever:

$$\sum_{i=1}^{D(t)} W_i \leq \int_{s=0}^{t} Q(s)ds \leq \sum_{i=1}^{A(t)} W_i$$

A fila média é dada por $L^{(t)} = (1/t)\int_{s=0}^{t} Q(s)ds$ e as esperas médias por $W^{[A(t)]} = (1/A(t))\sum_{i=1}^{A(t)} W_i$ e $W^{[D(t)]} = (1/D(t))\sum_{i=1}^{D(t)} W_i$ para usuários que chegam antes de t e para usuários que deixam a fila antes de $t$, respectivamente, podendo-se escrever:

$$D(t)W^{[D(t)]} \leq tL^{(t)} \leq A(t)W^{[A(t)]}$$

ou:

$$(1/t)D(t)W^{[D(t)]} \leq L^{(t)} \leq (1/t)A(t)W^{[A(t)]}$$

Essa relação tende para a fórmula de Little $L = \lambda W$ se:

$$(1/t)A(t) \to \lambda, (1/t)D(t) \to \lambda, L^{(t)} \to L \text{ para } t \to \infty$$

e $W^{[i]} \to W$ para $i \to \infty$.

Observe-se que essa derivação da fórmula de Little (1961) requer apenas a existência de limites de certas características operacionais do sistema, hipóteses que são normalmente utilizadas nos estudos de sistemas de filas no estado estacionário, não incluindo qualquer hipótese relativa à disciplina de atendimento.

Devido à dificuldade para se apresentar, de forma analítica, a generalização das equações de equilíbrio da fila, para o caso com $r$ classes de prioridades de chamados e com tamanho de fila limitado em $n = m - N$, em que, conforme anteriormente, m é o número de usuários no sistema e $N$ é o número de servidores, um procedimento que generaliza a geração dessas equações para $r$ classes de prioridades e capacidade de fila $s$ pode ser visto em Souza (2010).

## 3. Exemplo ilustrativo do modelo hipercubo com prioridades

Um exemplo de caráter ilustrativo considerando a disciplina de fila com prioridade foi desenvolvido considerando $r = 3$ classes de usuários: *a*, chamados de emergência; *b*, chamados de urgência moderada; *c*, chamados de urgência leve. O objetivo é representar um sistema simplificado, mas com as mesmas características básicas do SAMU-RP. Para a aplicação do modelo hipercubo é necessário considerar algumas premissas, escolhidas de forma a tornar o exemplo ilustrativo o mais parecido possível do SAMU-RP:

- A região é particionada em três átomos geográficos ($N_A = 3$), como mostrado na Figura 3;
- O sistema tem $N = 3$ servidores, localizados da seguinte forma: os servidores 1 (VSA, Veículo de Suporte Avançado) e 3 (VSB, Veículo de Suporte Básico) estão localizados no átomo 1; o servidor 2 (VSB) está localizado no átomo 2;
- O sistema possui $r = 3$ classes de usuários;
- Cada átomo geográfico $j$ ($j = 1,..., N_A$) foi subdividido em três camadas ou subátomos (não geográficos), que representam as classes de prioridades. Por exemplo, o átomo 1 é dividido nos subátomos 1*a*, 1*b* e 1*c*, representando as classes *a*, *b* e *c* no átomo 1, conforme Figura 3. Conforme anteriormente, seja $D = \{a, b, c\}$ o conjunto dessas classes de usuários. Pode-se notar que, dessa maneira, o sistema, originalmente com $N_A = 3$ átomos geográficos, passa

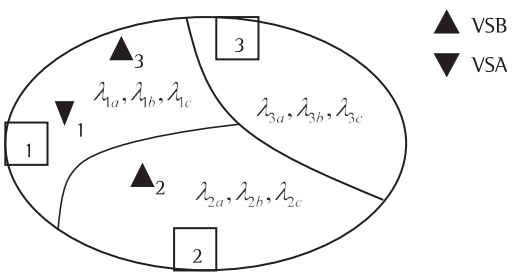


**Figura 3.** Átomos geográficos.

a ser analisado como um sistema alternativo com $N_A \times / D / = 3 \times 3 = 9$ subátomos. Admite-se que em cada subátomo os chamados chegam de acordo com o processo de Poisson. As taxas de chegada podem ser escritas em função do átomo, da classe ou do sistema:

$\lambda_j = \sum_{k \in D} \lambda_{jk}$ é a taxa de chegada de usuários do átomo *j*.

$\lambda_k = \sum_{j=1}^{N_A} \lambda_{jk}$ é a taxa de chegada de usuários da classe *k* no sistema

$\lambda = \sum_{j=1}^{N_A} \lambda_j = \sum_{j=1}^{N_A} \sum_{k \in D} \lambda_{jk}$; a taxa total de chegada no sistema é a soma dos $N_A$ átomos geográficos, ou a soma de todos os $N_A \times | D |$ subátomos do sistema.

A subdivisão dos átomos em subátomos para incorporar prioridades no modelo não invalida a hipótese do modelo hipercubo de que a área deve ser dividida em átomos (agora em subátomos) geográficos. Essa abordagem já foi utilizada em Takeda, Widmer e Morabito (2007) e Souza et al. (2010) para modelar as classes de usuários do SAMU-Campinas e SAMU-RP, respectivamente, no modelo hipercubo.

- Admite-se que o tempo de serviço segue uma distribuição exponencial para cada servidor *i* com taxa de serviço $\mu_i$, $i$ = 1, 2, 3, de forma que a taxa de serviço total do sistema seja $\mu = \sum_{i=1}^{N} \mu_i$;
- *m* é o número de usuários no sistema; e
- A lista de preferências de despacho encontra-se na Tabela 1. Os servidores podem viajar a qualquer subátomo, sendo que, para cada chamada, somente um servidor é enviado. Cada subátomo (1*a*, 2*a*, ..., 3*c*) possui um servidor primário e dois servidores *backup*. O servidor 1 (VSA) é enviado para atender a chamados de prioridade *b* e *c* somente se os servidores 2 e 3 (VSBs) estiverem ocupados. Assim, na lista de preferências da Tabela 1, o servidor 1 é o último a ser escolhido para ser enviado no atendimento a subátomos do tipo *b* e *c*, enquanto que nos subátomos do tipo *a* ele é o primeiro a ser escolhido. A Tabela 1 foi construída considerando o menor tempo médio de viagem de um servidor para cada subátomo. Todos os tempos médios de viagem do servidor entre dois subátomos estão na Tabela 2.

Para a resolução do modelo hipercubo é necessário definir os parâmetros do modelo:

- A matriz de preferência de despachos de ambulâncias para cada subátomo do sistema da Tabela 1;
- A matriz dos tempos médios de viagem $\tau_{jk,il}$ entre subátomos *jk* e *il*, da Tabela 2. Os tempos de viagem dessa matriz foram escolhidos arbitrariamente nos átomos 1 e 2. Como não há servidores localizados no átomo 3, os tempos de viagem dos subátomos 3*k* a qualquer outro subátomo são iguais a zero;
- A matriz de probabilidades da localização do servidor *i* em cada subátomo *jk* do sistema quando o servidor está desocupado, $l_{i,jk}$, da Tabela 3. Os servidores 1 (VSA) e 3 (VSB) estão localizados no átomo 1. O servidor 2 (VSBs) está localizado no átomo 2. Todos os servidores atendem aos chamados das classes *a*, *b* e *c* na mesma proporção;
- A matriz $t_{i,jk}$ pode ser obtida multiplicando a matriz *l* pela matriz $\tau$, $t_{i,jk} = \sum_{j=1}^{N_A} \sum_{k \in D} l_{i,jk} \tau_{jk,jl}$ (LARSON; ODONI, 2007).

O modelo utiliza disciplina de prioridade e admite fila com capacidade finita de, no máximo, três usuários em espera. Caso ocorra um chamado quando

**Tabela 1.** Lista de preferências de despacho de ambulâncias para cada subátomo do sistema.

| Subátomo | 1º | 2º | 3º |
|---|---|---|---|
| 1a | 1 | 3 | 2 |
| 1b | 3 | 2 | 1 |
| 1c | 3 | 2 | 1 |
| 2a | 1 | 2 | 3 |
| 2b | 2 | 3 | 1 |
| 2c | 2 | 3 | 1 |
| 3a | 1 | 2 | 3 |
| 3b | 2 | 3 | 1 |
| 3c | 2 | 3 | 1 |

**Tabela 2.** Matriz dos tempos médios de viagem entre os subátomos, $\tau_{il,jk}$(minutos) (*).

| | 1a | 1b | 1c | 2a | 2b | 2c | 3a | 3b | 3c |
|---|---|---|---|---|---|---|---|---|---|
| 1a | 10 | 10 | 10 | 12 | 12 | 12 | 15 | 15 | 15 |
| 1b | 10 | 10 | 10 | 12 | 12 | 12 | 15 | 15 | 15 |
| 1c | 10 | 10 | 10 | 12 | 12 | 12 | 15 | 15 | 15 |
| 2a | 13 | 13 | 13 | 10 | 10 | 10 | 14 | 14 | 14 |
| 2b | 13 | 13 | 13 | 10 | 10 | 10 | 14 | 14 | 14 |
| 2c | 13 | 13 | 13 | 10 | 10 | 10 | 14 | 14 | 14 |
| 3a | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3b | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3c | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

(*) Dados arbitrados

**Tabela 3.** Matriz de probabilidades da localização dos servidores em cada subátomo do sistema quando os servidores estão desocupados, $l_{i,jk.}$

| | 1a | 1b | 1c | 2a | 2b | 2c | 3a | 3b | 3c |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1/3 | 1/3 | 1/3 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 1/3 | 1/3 | 1/3 | 0 | 0 | 0 |
| 3 | 1/3 | 1/3 | 1/3 | 0 | 0 | 0 | 0 | 0 | 0 |

a fila já tem três usuários esperando, esse chamado será considerado uma perda para o sistema. Assim, a probabilidade de perda é mais uma medida de desempenho importante a ser considerada. O tempo médio de resposta é composto do tempo médio de viagem, o tempo médio de *setup* e o tempo médio de espera na fila. As tabelas a seguir apresentam o tempo de resposta sem considerar o tempo de *setup*, que pode ser considerado como cerca de 2 minutos. As taxas de chegada do usuário e as taxas de serviço no sistema são mostradas nas Tabelas 4 e 5, respectivamente.

Além disso, um modelo de simulação desse exemplo ilustrativo, que também considera explicitamente a prioridade na fila, foi desenvolvido e implementado no *software* Arena 7.0 com o objetivo de validar o modelo hipercubo estendido para analisar prioridades na fila de espera. Nesse modelo assumimos as mesmas características de política de despacho descritas anteriormente e utilizamos a lista de preferência da Tabela 1. A simulação considerada é contínua supondo que o SAMU é um sistema que trabalha continuamente (i. e., 24 horas, e 7 dias por semana). Assim sendo, aplicamos um procedimento de análise gráfica da média móvel (usando ferramentas do *software* Arena) para determinar o período transiente (*warm-up*). Para mais detalhes sobre esse procedimento e período transiente, o leitor pode consultar Banks (1998) e Kelton, Sadowski e Sadowski (2002). Na simulação desse exemplo ilustrativo, o tempo de *warm-up* utilizado foi de 125 dias. Utilizamos também outras ferramentas estatísticas do *software* Arena para determinar o tamanho da simulação, baseado na correlação entre diferentes lotes de observações das medidas de desempenho (para mais detalhes, veja e.g. KELTON; SADOWSKI; SADOWSKI, 2002; IANNONI; MORABITO, 2006).

Algumas estatísticas obtidas com os resultados de simulação foram comparadas com os resultados do modelo hipercubo. Entre estas: tempos de viagem, taxa de ocupação de cada ambulância (carga de trabalho), tempo de espera em fila para todas as chamadas e somente para chamadas que realmente esperam em fila, tempo de espera em fila para cada classe de usuário, frequências de despacho das ambulâncias e tempo médio no sistema. Os resultados dos dois modelos devem convergir, a menos de um erro amostral.

A Tabela 6 mostra os resultados dos tempos médios de espera na fila (calculados a partir da fórmula de Little, discutida na seção 2) para todos os chamados do sistema ($W_q$) e para as classes *a*, *b* e *c* ($W_{qa}$ , $W_{qb}$ e $W_{qc}$), calculados a partir da fórmula de Little discutida na seção 2 no caso de servidores heterogêneos, e pela simulação. Note que os desvios foram pequenos em todas as situações; o maior desvio encontrado foi de 2,36%.

## 4. Experimentos usando dados do SAMU-RP

Conforme descrito no estudo de caso realizado no SAMU-RP em Souza (2010), no SAMU-RP são atendidas três classes de usuários: chamados graves (emergências), atendidos prioritariamente pelo VSA (Veículo de Suporte Avançado); chamados moderados (urgências 1) e chamados leves (urgências 2), atendidos por um VSB (Veículo de Suporte Básico). O SAMU-RP também faz remoção de pacientes caracterizado pelo transporte de pacientes entre hospitais, de casa para hospital ou vice-versa. Esses atendimentos são na maioria agendados, de acordo com as operações desse sistema e dos hospitais envolvidos. O atendimento a ocorrências leves e remoção faz com que a demanda pelo serviço aumente significativamente.

No SAMU-RP, não há um limite preestabelecido para o tamanho da fila de chamados. Os chamados, em geral, não são transferidos para outro sistema nos casos de congestionamento. As ambulâncias estão descentralizadas em cinco bases e o VSA não atende a chamados de prioridade moderada e leve. Para fins de modelagem, a fila foi limitada em cinco usuários e o VSA atende a qualquer chamado, de forma que sempre será o último a ser escolhido caso o chamado seja de prioridade diferente de *a*. Essas aproximações mostraram-se viáveis para a análise do SAMU-RP no estudo de caso realizado em Souza (2010) e

**Tabela 4.** Taxas de chegada (chamadas por hora) dos átomos (última linha - $\lambda_j$), das classes (última coluna - $\lambda_k$) e dos subátomos (células internas - $\lambda_{jk}$).

| Classe | Origem (átomo) | | | Total |
|---|---|---|---|---|
| | 1 | 2 | 3 | |
| *a* | 0,0213 | 0,0426 | 0,0213 | 0,0851 |
| *b* | 0,2532 | 0,1115 | 0,1003 | 0,4651 |
| *c* | 1,0059 | 1,3795 | 1,1774 | 3,5627 |
| Total | 1,2804 | 1,5335 | 1,2989 | 4,1128 |

**Tabela 5.** Taxas de serviço $\mu_i$ (hora) dos três servidores.

| Servidores | 1 | 2 | 3 |
|---|---|---|---|
| Taxa de serviço | 1,5068 | 2,2874 | 2,5164 |

**Tabela 6.** Tempo médio de espera na fila $W_q$, $W_{qa}$, $W_{qb}$ e $W_{qc}$.

| | Servidores heterogêneos (minutos) | Simulação (minutos) | Desvio (minutos) | Desvio (%) |
|---|---|---|---|---|
| $W_q$ | 5,59 | 5,53 | 0,06 | 1,11 |
| $W_{qa}$ | 3,27 | 3,19 | 0,08 | 2,36 |
| $W_{qb}$ | 3,45 | 3,41 | 0,04 | 1,31 |
| $W_{qc}$ | 5,92 | 5,86 | 0,07 | 1,12 |

Souza et al. (2010), uma vez que a probabilidade de haver mais que cinco usuários em fila é bem pequena (menor que $10^{-4}$) e a maior frequência com que o VSA atende chamados diferentes de *a* também é bem pequena (0,0856) no período da manhã.

O nível de serviço a ser oferecido e a configuração dos SAMUs devem ser escolhidos de acordo com os tipos de atendimentos mais solicitados, considerando, por exemplo, configurações geográficas do município, características da demanda local e regiões com diferentes índices de demanda ao longo do dia. Mais detalhes sobre o SAMU-RP podem ser encontrados em Souza (2010).

No presente estudo foram analisados dois cenários alternativos com dados baseados no SAMU-RP. No primeiro cenário analisa-se o sistema considerando os chamados de remoção, enquanto que no segundo cenário avalia-se o aumento de demanda no sistema nas três classes de usuários. Além disso, propomos analisar subperíodos do dia que podem ter diferentes características temporais e geográficas importantes de serem consideradas. Utilizamos uma abordagem baseada no modelo hipercubo para um sistema em equilíbrio, para analisar cada período do dia de operação do SAMU de forma independente. Por meio dessa abordagem é possível analisar o sistema a partir das medidas de desempenho obtidas pelo modelo para cada subperíodo do dia e também analisar cada período crítico do dia, possibilitando melhorar o atendimento em cada período estudado do ponto de vista das medidas de desempenho internas e externas do sistema.

## 4.1. *Cenário 1 – Considerando os chamados de remoção*

Conforme Souza (2010), todos os VSBs do SAMU-RP podem fazer atendimentos de remoção. O objetivo dessa análise é ilustrar o impacto dos atendimentos de remoção, avaliando como essa sobrecarga de trabalho nas ambulâncias pode afetar o desempenho do sistema. Na falta de dados específicos disponíveis dos atendimentos de remoção no SAMU-RP, esse cenário considera algumas hipóteses simplificadoras: os chamados de remoção chegam aleatoriamente, de acordo com um processo de Poisson; os chamados são atendidos assim que uma ambulância fica desocupada, ou seja, sem política de agendamento programado; as remoções são classificadas como chamados de prioridade *c*; os tempos de atendimento desses chamados são considerados exponencialmente distribuídos. Na prática, o SAMU-RP realiza a maior parte das transferências de pacientes de forma programada, levando em conta a disponibilidade das ambulâncias e, às vezes, transportando mais de um paciente na mesma viagem da ambulância.

Dessa maneira, a proporção de chamados do tipo remoção (1,6667 chamado/hora) foi simplesmente acrescentada nos átomos *c* (1*c*, 2*c*, 3*c*, 4*c* e 5*c*). A Tabela 7 corresponde aos dados do estudo de caso em Souza (2010) com as modificações nos átomos *c*. Esse acréscimo aumentou consideravelmente a demanda no sistema, uma vez que, apesar da baixa prioridade, o volume desse tipo de atendimento é bastante representativo, da ordem de 60% de todos os atendimentos do sistema, segundo Souza (2010). Além das taxas de chegada dos chamados nos átomos *c*, os tempos médios de atendimento das ambulâncias também mudam na presença de atendimentos de remoção. Porém, como não se dispõem de dados dos atendimentos de remoção, por simplicidade, neste cenário consideraram-se os mesmos tempos médios de atendimento reportados para o cenário original em Souza (2010). A Figura 4 apresenta as taxas de utilização de cada ambulância (*workloads*) para o cenário 1. Os resultados mostram que as *workloads* são substancialmente altas (em torno de 80% para os períodos da manhã e tarde e em torno de 70% para o período da noite), se considerarmos que se trata de um serviço emergencial.

O modelo de simulação mencionado na seção 3 foi aplicado para validar os resultados do modelo hipercubo estendido. Nos dois modelos foram consideradas distribuições exponenciais para os intervalos de tempo entre as chegadas e os tempos de serviços. Além disso, a política de despacho foi

**Tabela 7.** Taxas médias de chegada dos chamados para cada subátomo em cada período para o cenário 1.

| Subátomos | | $\lambda_{jk}$ (Chamados com remoção) | | |
|---|---|---|---|---|
| | | Manhã | Tarde | Noite |
| 1 | 1*a* | 0,116381 | 0,127671 | 0,025582 |
| 2 | 1*b* | 0,675009 | 0,638355 | 0,255815 |
| 3 | 1*c* | 2,015813 | 1,964569 | 1,820159 |
| 4 | 2*a* | 0,116381 | 0,085114 | 0,025582 |
| 5 | 2*b* | 0,558628 | 0,659634 | 0,562793 |
| 6 | 2*c* | 2,318403 | 2,283747 | 1,922485 |
| 7 | 3*a* | 0,116381 | 0,021279 | 0,025582 |
| 8 | 3*b* | 0,558628 | 0,553241 | 0,486049 |
| 9 | 3*c* | 1,969260 | 1,985848 | 1,768996 |
| 10 | 4*a* | 0,023276 | 0,042557 | 0,000000 |
| 11 | 4*b* | 0,256038 | 0,191507 | 0,153489 |
| 12 | 4*c* | 1,806327 | 1,879455 | 1,871322 |
| 13 | 5*a* | 0,046552 | 0,021279 | 0,000000 |
| 14 | 5*b* | 0,581904 | 1,170318 | 0,409304 |
| 15 | 5*c* | 1,945984 | 2,177354 | 1,922485 |
| Total | | 13,10497 | 13,80193 | 11,24964 |

modificada nos dois modelos para melhor representar o SAMU-RP, em que não há uma lista de preferência fixa para todos os átomos. Dessa forma, quando uma chamada chega ao sistema, verifica se:

- Se a chamada é do tipo *a*, e se o servidor VSA está disponível, esse servidor passa a ser ocupado (é despachado). Se esse servidor estiver ocupado, verifica-se a disponibilidade dos servidores preferenciais do átomo (origem dessa chamada). Se houver dois servidores preferenciais no átomo e os dois estiverem disponíveis, escolha-se um servidor de forma aleatória. Ou se apenas um servidor estiver ocupado, este é despachado. Se estes estiverem ocupados, verifica-se a disponibilidade dos demais servidores (localizados em outros átomos), e escolha de forma aleatória entre os servidores *backup* disponíveis;
- Se a chamada é do tipo *b* ou *c*, verifica se a disponibilidade dos servidores preferenciais do átomo (podem ser 1 ou 2 servidores, dependendo do átomo). Se os dois servidores estiverem disponíveis, escolha-se de forma aleatória entre os dois. Se somente um estiver disponível, este é despachado (se torna ocupado). Se os servidores preferenciais estiverem ocupados, escolha-se de forma aleatória entre os demais servidores (fora do átomo) disponíveis. Se todos os servidores VSB estiverem ocupados, a chamada é atendida pela VSA, caso esta esteja livre; e
- Caso todos os servidores estiverem ocupados, a chamada entra em fila se esta tiver menos que cinco chamadas em espera. Caso a fila esteja cheia, a chamada é perdida para o sistema.

Para considerar no modelo hipercubo uma lista de preferência despacho em que um átomo tem mais de um servidor preferencial (i. e., casos de desempate de prioridade entre ambulâncias de mesmo local), considerou-se um número suficientemente grande de listas de despacho geradas aleatoriamente, representando, dessa forma, as possíveis chances dos servidores primários (e/ou *backup*) de cada átomo serem enviados para atender um chamado em cada cenário investigado, conforme em Burwell, Jarvis e McKnew (1993) e Takeda, Widmer e Morabito (2007). O tempo de *warm-up* considerado modelo de simulação foi de 125 horas, conforme seção 3.

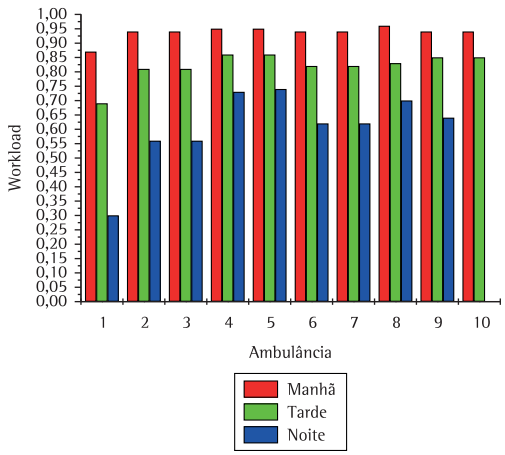


**Figura 4.** *Workloads* das ambulâncias em cada período para o cenário 1.

Para adaptar o modelo hipercubo, descrito na seção 2, o método escolhido para realizar a política de despacho descrita anteriormente foi da geração aleatória da política de despacho dos servidores, sugerido por Burwell, Jarvis e McKnew (1993). Nos átomos Central, Norte, Sul e Oeste há duas ambulâncias VSBs. Elas são escolhidas preferencialmente e, com a mesma chance, para chamados *b* ou *c*, ocupando a 1ª e 2ª preferências. Caso as duas ambulâncias estejam ocupadas, qualquer ambulância VSB pode atender ao chamado. Assim, é feito um sorteio para a escolha das próximas preferências, quando todas as ambulâncias VSBs restantes têm a mesma chance de serem escolhidas em todas as posições. O VSA atende a esses chamados somente se todas as ambulâncias VSBs estiverem ocupadas. Na região Leste, há apenas uma ambulância, de forma que ela é sempre a primeira escolha para chamados *b* ou *c* e o procedimento para a escolha das próximas posições é semelhante aos átomos com 2 ambulâncias.

A Tabela 8 mostra os tempos médios de espera em fila no sistema e para os átomos *a*, *b* e *c* para o período da tarde, obtidos com modelo hipercubo e simulação. Note que os desvios do modelo com a simulação são pequenos, todos menores que 5%. Nesse cenário, a probabilidade de todos os servidores estarem ocupados ($P_{1111111111}$) é de 0,1000 e a probabilidade de fila é de 0,3238.

Conforme reportado em Souza (2010), no cenário original do SAMU-RP o tempo médio de espera na fila é de 0,02 minutos, de forma que os tempos médios de viagem coincidem com os tempos médios de resposta. A partir da análise desse cenário pôde-se observar que, considerando atendimentos de remoção, há um grande aumento no tempo médio de espera na fila em todos os tipos de chamados e os tempos de resposta são afetados diretamente com esse aumento.

**Tabela 8.** Comparação dos tempos médios de espera em fila de cada classe de usuários obtidos pelo modelo e via simulação.

| Tempo médio de espera em fila (min.) | | Sistema | *a* | *B* | *c* |
|---|---|---|---|---|---|
| Modelo | | 3,5340 | 1,4000 | 1,6000 | 4,2000 |
| Simulação | | 3,5016 | 1,3854 | 1,5918 | 4,1562 |
| Desvio | Minutos | 0,0324 | 0,0146 | 0,0082 | 0,0438 |
| | % | 0,9 | 1,1 | 0,5 | 1,1 |

**Tabela 9.** *Workloads* das ambulâncias no cenário 2 para diferentes aumentos de demanda no período da manhã.

| Ambulância | *Workload* | | | | |
|---|---|---|---|---|---|
| | Sistema original | +10% | +25% | +50% | +150% |
| 1 | 0,1600 | 0,1906 | 0,3377 | 0,3595 | 0,8250 |
| 2 | 0,4150 | 0,4587 | 0,5325 | 0,6188 | 0,9020 |
| 3 | 0,4187 | 0,4586 | 0,5317 | 0,6201 | 0,9015 |
| 4 | 0,5290 | 0,5677 | 0,6294 | 0,7030 | 0,9264 |
| 5 | 0,5280 | 0,5678 | 0,6341 | 0,7034 | 0,9271 |
| 6 | 0,4265 | 0,4732 | 0,5511 | 0,6317 | 0,9090 |
| 7 | 0,4312 | 0,4691 | 0,5520 | 0,6302 | 0,9091 |
| 8 | 0,4590 | 0,5016 | 0,5868 | 0,6674 | 0,9257 |
| 9 | 0,3943 | 0,4355 | 0,5496 | 0,6022 | 0,9020 |
| 10 | 0,3902 | 0,4340 | 0,5480 | 0,6036 | 0,9012 |
| VSA | 0,1600 | 0,1906 | 0,3377 | 0,3595 | 0,8250 |
| VSB | 0,4435 | 0,4851 | 0,5684 | 0,6423 | 0,9116 |
| Total | 0,4152 | 0,4557 | 0,5453 | 0,6140 | 0,9029 |

Os resultados desse cenário indicam o quanto os atendimentos de remoção podem piorar o desempenho do sistema, do ponto de vista do usuário.

### 4.2. *Cenário 2 – Aumento de demanda*

No cenário 2, analisa-se o impacto do aumento de demanda no período da manhã do cenário original estudado em Souza (2010), resultando em maiores taxas de utilização do sistema em todas as classes de usuários. O modelo hipercubo, descrito na seção 2, foi utilizado para fazer a análise desse cenário. Como os desvios do modelo com relação à simulação foram pequenos na análise do cenário 1, a simulação não foi realizada para avaliação desse cenário. O objetivo é analisar o comportamento das medidas de desempenho do sistema sob dois pontos de vista: do gerente do sistema, e para isso foi escolhida uma medida interna, o *workload*; e do usuário do sistema, e para isso foi escolhida uma medida externa, o tempo médio de espera na fila. A Tabela 9 mostra o aumento do *workload* com o aumento da demanda, a utilização média das ambulâncias varia de 46% (com aumento de demanda de 10%) até chegar a 90% (com aumento de demanda de 150%).

A Tabela 10 mostra os resultados dos tempos de espera na fila quando a demanda do sistema original do SAMU-RP tem aumentos de 10% até 150%. Note que o tempo médio de espera na fila aumenta significativamente em relação ao sistema original. Os resultados mostram que seria necessário aumentar a demanda em 150% de chamados das classes *a*, *b* e *c* para que as medidas de desempenho analisadas ficassem próximas das medidas de desempenho do

**Tabela 10.** Tempos médios de espera na fila das classes de usuários no cenário 2 para diferentes aumentos de demanda no período da manhã.

| Demanda | Tempo médio de espera em fila (min.) | | | |
|---|---|---|---|---|
| | Sistema | *a* | *b* | *C* |
| Sistema original | 0,0849 | 0,0541 | 0,0737 | 0,1124 |
| +10% | 0,1629 | 0,0989 | 0,1390 | 0,2219 |
| +25% | 0,5391 | 0,3007 | 0,4591 | 0,7914 |
| +50% | 1,1013 | 0,5519 | 0,8700 | 1,6575 |
| +150% | 9,0667 | 3,1763 | 5,7846 | 16,1857 |

cenário 1, que considerou o impacto do atendimento de remoção no sistema sem agendamento programado.

## 5. Conclusões

Este trabalho propõe uma extensão do modelo hipercubo de filas espacialmente distribuídas para analisar a configuração de um SAMU. Em sistemas de atendimento emergencial mais congestionados, a formação de fila e a disciplina de prioridade no atendimento da fila são fatores importantes a serem considerados. Esses sistemas diferenciam os usuários em classes e atribuem prioridades no atendimento de acordo com algum critério, por exemplo, a urgência do chamado do usuário. Para representar mais adequadamente sistemas com essas características, neste estudo o modelo hipercubo foi estendido para considerar explicitamente políticas de prioridade no atendimento da fila dos usuários. Foram desenvolvidas as equações de equilíbrio de cada estado da cauda (fila) do modelo hipercubo, juntamente com medidas de desempenho considerando prioridade na fila. Por meio dessa extensão, pôde-se obter medidas de desempenho por classe (por exemplo, os tempos médios de espera na fila e os tempos médios de resposta aos usuários de cada classe), e não apenas medidas para todos os chamados agrupados, como no modelo hipercubo clássico.

O cenário 1 foi desenvolvido para analisar o impacto do atendimento de remoção de pacientes no sistema. Para isso, as taxas de chegada dos chamados da classe *c* foram aumentadas para avaliar o impacto de se atender remoções de pacientes. Convém salientar que isso ainda não é um cenário bem representativo dos atendimentos de remoção na prática do SAMU-RP, uma vez que não foram coletados dados sobre os atendimentos de remoção, nem foram fornecidas informações suficientes para a estimativa das distribuições das chegadas e dos serviços desses atendimentos e, ainda, como eles influenciam os atendimentos das classes *a*, *b* e *c*. Além disso, as remoções podem ser agendadas previamente e os atendimentos podem ser agrupados (por exemplo,

uma ambulância pode fazer duas ou três remoções antes de voltar para a base), descaracterizando a aleatoriedade dos chamados e seus atendimentos, um a um (e não em lote), por uma ambulância.

No cenário 2, o objetivo foi avaliar o impacto do aumento de demanda no período da manhã dos chamados das classes *a*, *b* e *c*, no período mais congestionado do dia (o período da manhã). Assim, a demanda dos chamados dessas classes foi simplesmente aumentada para 10%, 25%, 50% e 150%. Verificou-se que as medidas de desempenho são consistentes para aumentos de até 25% de demanda. Com aumentos de 50% ou mais, os tempos médios de resposta aumentaram significativamente, se comparados com os tempos médios de viagem. Quando a demanda dos chamados dessas classes aumenta para mais de 100%, as medidas de desempenho ficam próximas das medidas de desempenho obtidas a partir do cenário 1 (atendimento de remoções sem agendamento programado).

Durante o desenvolvimento desse estudo surgiram algumas perspectivas interessantes para pesquisas futuras. Algumas possibilidades seriam conduzir outras análises de sensibilidade além de alterações na demanda do sistema e estudar outros cenários alternativos explorando outras mudanças de configuração do sistema. Também seria interessante analisar o sistema considerando os atendimentos de remoção como uma quarta classe de usuários, uma vez que esses atendimentos podem representar uma parcela relevante do total de atendimentos do sistema. Porém, um questionamento com relação a isso é se essas transferências deveriam mesmo ser realizadas pelo sistema SAMU, que é um sistema de atendimento eminentemente emergencial.

Outra linha de pesquisa interessante refere-se a situações em que a ambulância, logo após terminar um atendimento, é imediatamente redespachada para outro atendimento a partir do local em que se encontra, isto é, sem que já tenha voltado para o local da base do sistema. Assim, seria interessante desenvolver abordagens de redespacho no modelo hipercubo com prioridade de fila para também considerar essas situações. Como no SAMU-RP o VSA não atende a chamados *b* e *c*, também poder-se-ia estudar como modificar o modelo hipercubo com prioridade na fila para considerar *backup* parcial e adaptar apropriadamente suas medidas de desempenho para um sistema com essas características. Outra pesquisa futura seria determinar o número mínimo de ambulâncias necessárias no sistema em diferentes períodos a fim de manter uma ou mais medidas de desempenho em um nível desejado. Uma alternativa para essas análises seria investigar a possibilidade de incorporar o modelo hipercubo em problemas dinâmicos, considerando a realocação e o reposicionamento das ambulâncias ao longo do dia.

## Referências

ATKINSON, J. B. et al. Heuristic methods for the analysis of a queuing system describing emergency medical service deployed along a highway. *Cybernetics and Systems Analysis*, v. 42, n. 3, p. 379-391, 2006. http://dx.doi.org/10.1007/s10559-006-0075-6

ATKINSON, J. B. et al. A hypercube queueing loss model with customer-dependent service rates. *European Journal of Operational Research*, v. 191, p. 223-239, 2008. http://dx.doi.org/10.1016/j.ejor.2007.08.014

BANKS, J. *Handbook of Simulation*. Atlanta: John Wiley & Sons, 1998. p. 3-389. http://dx.doi.org/10.1002/9780470172445

BRANDEAU, M.; LARSON, R. C. Extending and applying the hypercube queueing model to deploy ambulances in Boston. In: SWERSEY, A. J.; INGNALL, E. J. (Eds.). *Delivery of Urban Services*. Elsevier, 1986. p. 121-153. (TIMS Studies in the Management Science, n. 22).

BOFFEY, B.; GALVÃO, R. D.; ESPEJO, L. G. A. A review of congestion models in the location of facilities with immobile servers. *European Journal of Operational Research*, v. 178, p. 643-662, 2007. http://dx.doi.org/10.1016/j.ejor.2006.04.044

BURWELL, T. H.; JARVIS, J. P.; McKNEW, M. A. Modeling co-located servers and dispatch ties in the hypercube model. *Computers & Operations Research*, v. 202, n. 2, p. 113-119, 1993. http://dx.doi.org/10.1016/0305-0548(93)90067-S

CHELST, K. R.; BARLACH, Z. Multiple unit dispatches in emergency services: models to estimate system performance. *Management Science*, v. 272, n. 12, p. 1390-1409, 1981. http://dx.doi.org/10.1287/mnsc.27.12.1390

CHIYOSHI, F.; GALVÃO, R. D.; MORABITO, R. O uso do modelo hipercubo na solução de problemas de localização probabilísticos. *Gestão & Produção*, v. 72, n. 2, p. 146-174, 2000.

CHIYOSHI, F.; GALVÃO, R. D.; MORABITO, R. Modelo hipercubo: análise e resultados para o caso de servidores não-homogêneos. *Pesquisa Operacional*, v. 212, n. 2, p. 199-218, 2001.

GALVÃO, R. D.; MORABITO, R. Emergency service systems: The use of the hypercube queueing model in the solution of probabilistic location problems. *International Transactions in Operational Research*, v. 15, p. 525-549, 2008. http://dx.doi.org/10.1111/j.1475-3995.2008.00654.x

IANNONI, A.; MORABITO, R. A discrete simulation analysis of a logistics supply system. *Transportation Research Part E*, v. 42, p. 191-210, 2006. http://dx.doi.org/10.1016/j.tre.2004.10.002

IANNONI, A. P.; MORABITO, R. A multiple dispatch and partial backup hypercube queuing model to analyze emergency medical systems on highways. *Transportation Research E*, v. 432, n. 6, p. 755-771, 2007. http://dx.doi.org/10.1016/j.ejor.2008.02.003

IANNONI, A. P.; MORABITO, R.; SAYDAM, C. An optimization approach for ambulance location and the districting of the response segments on highways. *European Journal*

*of Operational Research*, v. 195, p. 528-542, 2009. http://dx.doi.org/10.1016/j.ejor.2008.02.003

KELTON, W. D.; SADOWSKI, R. P.; SADOWSKI, D. *A Simulation with Arena*. 2nd ed. New York: McGraw- Hill, 2002.

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

LARSON, R. C.; ODONI, A. R. *Urban Operations Research*. 2nd ed. Belmont: Dynamic Ideas, 2007. http://dx.doi.org/10.1016/0305-0548(74)90076-8

LITTLE, J. D. A proof for the queueing formula: L = .. *Operations Research*, v. 9, p. 383-387, 1961. http://dx.doi.org/10.1287/opre.9.3.383

MARIANOV, V.; RÍOS, M. A probabilistic quality of service constraint for a location model of switches in ATM communications networks. *Annals of Operations Research*, v. 96, p. 237-243, 2000. http://dx.doi.org/10.1023/A:1018955603355

MARIANOV, V.; SERRA, D. Probabilistic maximal covering location–allocation for congested system. J*ournal of Regional Science*, v. 38, p. 401-424, 1998. http://dx.doi.org/10.1111/0022-4146.00100

MARIANOV, V.; SERRA, D. Location-allocation of multiple-server service centers with constrained queues or waiting times. *Annals of Operations Research*, v. 111, p. 35-50, 2003. http://dx.doi.org/10.1023/A:1020989316737

MENDONÇA, F.; MORABITO, R. Analysing emergency medical service ambulance deployment on a Brazilian highway using the hypercube model. *Journal of the Operational Research Society*, v. 52, p. 261-270, 2001. http://dx.doi.org/10.1057/palgrave.jors.2601097

MORABITO, R.; CHIYOSHI, F.; GALVÃO, R. Non-homogeneous servers in emergency medical systems: practical applications using the hypercube queuing model. *Socio-Economic Planning Sciences*, v. 42, p. 255-270, 2008. http://dx.doi.org/10.1016/j.seps.2007.04.002

SACKS, S. R.; GRIEF, S. Orlando Police Department uses OR/MS methodology, new software to design patrol districts. Baltimore: OR/MS Today, 1994. p. 30-32.

SOUZA, R. M. *Análise da configuração de SAMU utilizando modelo hipercubo com prioridade na fila e múltiplas alternativas de localização de ambulâncias*. 2010. Tese (Doutorado em Engenharia de Produção)-Universidade Federal de São Carlos, São Carlos, 2010.

SOUZA, R. M. et al. Análise da configuração de SAMU utilizando múltiplas alternativas de localização de ambulâncias. *Gestão & Produção*, v. 12, n. 3, p. 333-345, 2010. http://dx.doi.org/10.1590/S0104-530X2005000300005

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

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, p. 727-741, 2007. http://dx.doi.org/10.1016/j.cor.2005.03.022

WHITT, W. A review of and extensions. *Queueing Systems*, v. 9, p. 235-268, 1991. http://dx.doi.org/10.1007/BF01158466

## Agradecimentos

Os autores agradecem os dois revisores anônimos pelos úteis comentários e sugestões e também ao SAMU-RP pela colaboração com esta pesquisa, que também contou com apoio do CNPq.

# Hypercube model extension for the analysis of emergencial medical systems with priority queue

## Abstract

In some emergency medical systems the service demand is high due to the treatment of patients in the range severe to mild. In these systems, may be queues formation and so the need to explicitly consider priority in care is extremely important. In this study we extend the hypercube model to explicitly consider priority queue. In order to verify the feasibility and applicability of this approach, we conducted a case study at Ribeirão Preto´s SAMU (SAMU-RP). We analyzed two alternative scenarios to examine two important issues: the impact of the removals and the effect of increased demand in the different classes of calls of the system. The focus is on the average response time to users, considered as an important performance measure of the system, especially for the high priority calls. The results show that the approach can be successfully used to analyze systems with priority queue.

## Keywords

Hypercube queuing model. Priority queue. SAMU. Emergency medical system.