# BACKUP PARCIAL NO ATENDIMENTO ÀS CLASSES DE CHAMADOS NO MODELO HIPERCUBO COM PRIORIDADE NA FILA

**Caio Vitor Beojone**
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Av. Eng. Luiz Edmundo C. Coube 14-01, Vargem Limpa – Bauru/ SP
beojone@hotmail.com

**Regiane Máximo de Souza**
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Av. Eng. Luiz Edmundo C. Coube 14-01, Vargem Limpa – Bauru/ SP
regiane@feb.unesp.br

RESUMO

Os Sistemas de Atendimento Emergenciais (SAE´s) são de grande importância para a sociedade, um SAE gerenciado com eficiência tanto do ponto de vista gerencial como dos usuários pode melhorar a qualidade de vida da população, seja na área da saúde, segurança etc. O modelo hipercubo vem se mostrando uma ferramenta importante a ser utilizada nesses sistemas a fim de avaliar seu desempenho e analisar cenários alternativos. Os SAE´s, em geral, possuem particularidades que devem ser incorporadas no modelo hipercubo original a fim de obter uma melhor acurácia da análise dos sistemas reais e, consequentemente, dos cenários alternativos. O objetivo desse trabalho é estender o modelo hipercubo para considerar prioridade na fila e backup parcial quanto às classes dos chamados sem restrição física, particularidades encontradas principalmente em SAMU´s recentemente estudados na literatura. Para melhor compreensão da extensão proposta, foi analisado um exemplo ilustrativo de acordo com as novas hipóteses do modelo.

**Palavras chave: Sistemas de Atendimento Emergenciais, modelo hipercubo.**

ABSTRACT

The Emergency Service Systems (SAE's) are of great importance for society, an SAE managed efficiently both point of view users and managers can improve people's quality of life, whether in health, security, etc. The hypercube model has proven to be an important tool to be used in these systems to assess their performance and analyze alternative scenarios. The SAE's generally possess characteristics that should be incorporated in the original hypercube model to obtain a better accuracy of the analysis of the real system, and hence the alternative scenarios. The aim of this paper is to extend the hypercube model to consider priority in the queue and partial backup for the class of the users and there is no physical restraint, features found in SAMU's recently studied in the literature. To better understand the proposed extension was considered one illustrative example in accordance with the new assumptions of the model.

**Keywords: Emergency Service Systems, hypercube model.**

## 1 INTRODUÇÃO

O modelo hipercubo tem sido utilizado recentemente para analisar diversos sistemas importantes de Atendimento Emergencial que aparecem na sociedade como saúde, segurança, manutenção, etc. Alguns desses sistemas são relacionados a área da saúde, como os SAMU´s, os SAE´s de rodovias, os corpos de bombeiros, as unidades de atendimentos emergências nos Pronto Socorros PS´s, entre outras. Na área de segurança, temos o serviço policial e de segurança privada, entre outros. Na área de manutenção emergencial tem-se exemplos em usinas Sucroalcooleiras, como em Rodriques (2014).

Em todos os casos, a resposta rápida a tal demanda é fundamental para minimizar possíveis sequelas decorrentes no quadro dos pacientes ou prejuízos importantes nos sistemas envolvidos. No Brasil, em 2007, mais de 35.000 pessoas perderam a vida vítimas de acidentes de trânsito (OPAS, 2008), porém acredita-se que esses números sejam ainda maiores. Só nas rodovias federais, ocorreram 128.456 acidentes, sendo 5.757 acidentes com mortes e 75.462 sem vítimas (DNIT, 2009). Em todo o mundo, o trânsito causa perda de vidas, mas os números brasileiros são alarmantes. Ainda em 2007, conforme a OMS (2009), o país ocupou o 5º lugar em mortes no trânsito no mundo.

Nos Sistemas de Atendimento de Emergência (SAE's), de forma geral, o tempo médio de resposta ao usuário é de fundamental importância, pois a demora no atendimento pode significar a vida ou a morte de uma pessoa. Devido às restrições orçamentárias, os SAE's não podem ter um grande número de pessoas e equipamentos, com mais servidores e tripulações. Assim, existe um compromisso (*trade-off*) evidente entre investimentos, custos operacionais e o nível de serviço oferecido aos usuários. É importante analisar estes sistemas considerando suas particularidades e seus recursos a fim de diminuir o tempo de resposta ao usuário.

Os Serviços de Atendimentos Emergenciais (SAE's) são caracterizados essencialmente por incertezas quanto a disponibilidade, localização, tempo de serviço dos servidores, demanda ao longo da região e tempo de resposta para atendimento aos usuários. Os sistemas de atendimento a emergências em saúde caracterizam um grande desafio para todas as nações, pois independentemente do tipo de urgência envolvida, somente com uma rigorosa organização é possível oferecer um serviço de boa qualidade (Souza, 2010). Nos últimos 50 anos, vem aumentando o interesse de pesquisadores em estudos de sistemas emergenciais (Simpson & Hancock, 2009).

Como os SAE´s são e serviços do tipo servidor-para-cliente (*server-to-costumer*), em que os servidores precisam se deslocar até o local da solicitação do usuário, a análise do seu funcionamento em geral, precisa levar em conta fatores probabilísticos na distribuição espacial e temporal dos chamados e servidores (Iannoni, 2005).

Nos sistemas servidor-para-cliente, os servidores viajam até o local onde acontece a emergência. Larson e Odoni (2007) citam como exemplos de sistemas servidor-para-cliente, além dos SAE's já citados, visitas em casa do serviço social e sistemas de entregas como pizzas e peças de carros. Nesses sistemas, os servidores são geralmente modelados como espacialmente distribuídos, com operação independente, com carga de trabalho diferente e com carga de trabalho variável, dependendo da localização do servidor (Galvão & Morabito, 2008; Chiyoshi et al., 2011).

Larson (1974) desenvolveu o modelo hipercubo de filas, um modelo descritivo aplicado a sistemas servidor-para-cliente que fornece as probabilidades de equilíbrio dos possíveis estados do sistema. O objetivo desse modelo é avaliar a configuração e estimar as medidas de desempenho relevantes para um sistema de atendimento de urgência, de forma a possibilitar um planejamento adequado e melhores níveis do serviço oferecido (Larson & Odoni, 2007). Esse modelo ainda permite ainda incorporar características específicas e particularidades de despacho de cada sistema emergencial analisado (Chiyoshi et al., 2011).

Suas medidas de desempenho 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 (referida neste estudo como átomo) 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 tem se mostrado eficiente e preciso para analisar SAE's, como foi analisado, por exemplo, nos Estados Unidos, em Chelst e Barlach (1981), Brandeau e Larson (1986), Burwell et al. (1993), Sacks e Grieff (1994), Swersey (1994) e Larson e Odoni (2007). No Brasil, alguns exemplos aparecem em Chiyoshi et al. (2000), Gonçalves et al. (1994), Gonçalves et al. (1995), Mendonça e Morabito (2000), Oliveira (2003), Chiyoshi et al. (2000, 2001), Costa et al. (2004), Figueiredo et al. (2005), Takeda et al. (2007), Iannoni (2005), Souza et al. (2013, 2014).

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 estado do sistema, possibilitando o cálculo das medidas de desempenho supracitadas.

Atualmente o modelo hipercubo foi adaptado para considerar sistemas com restrições físicas de atendimento, prioridade e *backup* parcial. Para esse tipo de sistema, a construção dos estados da fila precisa de um maior detalhamento, ou seja, além de separar de acordo com a prioridade é preciso separar de acordo com a origem do chamado (Rodrigues, 2014). Outra abordagem recentemente estudada em Iannoni *et. al.* (2015) foi considerar reserva de capacidade, este tipo de sistema está ligado a diferenças nas prioridades dos chamados. Levando em conta chamados mais graves, estes não podem esperar por atendimento, ou seja, precisam ser atendidos de imediato. Por outro lado, chamados mais simples, podem ter certo tempo de espera até ter um servidor a caminho do chamado. Isso acontece quando procura-se reservar a capacidade de servidores com relação a chamados mais graves. Para tanto, essa abordagem utiliza outra classificação para os estados de fila os quais levam em conta o tipo de urgência, o servidor livre (caso exista), e o número de chamados em espera. Em encontra-se a explicação detalhada dessa extensão, assim como um exemplo ilustrativo.

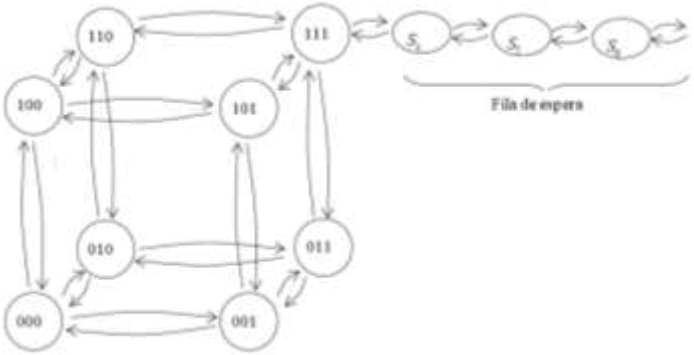
Nesse sentido, é importante adaptar o Modelo hipercubo de forma que ele se adeque aos diferentes sistemas com mais acurácia dos resultados. O objetivo desse trabalho é estender o modelo hipercubo para considerar prioridade na fila e *backup parcial* em sistemas em que não há restrição física, mas sim quanto às classes dos chamados.

## 2 MODELO HIPERCUBO CLÁSSICO

O modelo baseia-se na divisão da região atendida pelo sistema em átomos geográficos (regiões de demanda). Cada átomo é considerado uma fonte de chamados pontual e independente das demais e o atendimento a cada átomo é realizado por servidores que estão distribuídos na região. A localização dos servidores deve ser conhecida ou, caso contrário, estimada (Larson & Odoni, 2007). Se um servidor estiver ocupado, outros servidores poderão atender ao chamado, mesmo que estejam fora de sua região preferencial do chamado, prevalecendo a cooperação entre os servidores. Para considerar prioridade no despacho de ambulâncias quando o sistema não se encontra saturado, é realizada a estratégia de "*layering*" (Larson & Odoni, 2007) que consiste na divisão dos átomos geográficos em subátomos, conforme visto em Souza *et. al.* (2015).

A disponibilidade dos servidores é representada por meio do espaço de estados dos servidores. Um estado particular do sistema sem fila é dado pela lista dos servidores que estão livres ou ocupados. Considere um sistema com apenas $m = 3$ servidores e sejam {000}, {001}, {010}, ..., {111} os $2^3 = 8$ possíveis estados do sistema, em que os 0's e 1's indicam se cada um dos três servidores está livre ou ocupado, respectivamente. Por exemplo, o estado {011} representa o estado em que o servidor 1 está livre e os servidores 2 e 3 estão ocupados (note que {011} descreve o estado dos servidores da esquerda para a direita). Assim, o espaço de estados desse sistema com três servidores pode ser representado por um cubo; no caso de haverem mais de três servidores, temos um hipercubo. A Figura 1 ilustra o espaço de estados desse sistema com três servidores. O modelo hipercubo trata tanto sistemas em que não é permitida a formação de fila, como aqueles em que quando todos os servidores estão ocupados, os chamados que chegam esperam em uma fila por meio da qual os usuários são atendidos à medida que os servidores tornam-se desocupados segundo a disciplina FCFS. Os demais estados S4, S5, S6, ..., da Figura 1 representam os estados com 1, 2, 3, ..., usuários na fila de espera do sistema, respectivamente.

Figura 1 – Estados do sistema com três servidores.




Segundo Larson e Odoni (2007), existem nove hipóteses principais que devem ser verificadas para a aplicação do modelo hipercubo:

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

Algumas destas hipóteses podem ser alteradas, como, por exemplo, múltiplo despacho e backup parcial, como em Chelst e Barlach (1981), Mendonça e Morabito (2000), Iannoni (2005) e Iannoni et. al. (2008, 2009).

A seguir, é apresentado o modelo hipercubo por meio de um exemplo simples, resolvido em Chiyoshi *et. al* (2001). Considere um sistema de emergência operando em uma região representada por três átomos, utilizando política de despacho de preferência fixa, mostrada na Tabela 1.

Tabela 1 – Matriz de Preferências de despacho

| **Matriz de despachos** | | | |
|---|---|---|---|
| **Átomos** | **Preferências** | | |
| | **1º** | **2º** | **3º** |
| **1** | 1 | 2 | 3 |
| **2** | 2 | 3 | 1 |
| **3** | 3 | 1 | 2 |

A solução do modelo é dada pela construção das equações de equilíbrio do sistema, que são definidas supondo-se que o sistema atinja o equilíbrio. Para cada estado do sistema, o fluxo que entra neste estado deve ser igual ao fluxo que sai dele. Em um sistema não saturado, com capacidade de fila infinita, as

probabilidades de estado do modelo hipercubo são calculadas a partir das equações de balanço, construídas a partir dos oito possíveis estados, descritos anteriormente nesta mesma seção.
Quando o sistema está no estado {000} (sistema vazio), ele passa para o estado {100} quando ocorre um chamado com origem no átomo 1, com taxa de ocorrência $\boldsymbol{\lambda_1}$. O mesmo acontece com o estado {010}, com taxa $\boldsymbol{\lambda_2}$, e para o estado {001}, com taxa $\boldsymbol{\lambda_3}$. Dessa forma, a taxa total de transição do estado {000} para outros estados é $\boldsymbol{\lambda = \lambda_1 + \lambda_2 + \lambda_3}$.
No sentido contrário, o estado {000} pode ser alcançado a partir do estado {100} quando o servidor 1 termina o atendimento, com taxa $\mu_1$; da mesma forma, a partir do estado {010}, com taxa $\mu_2$; e a partir de {001}, com taxa $\mu_3$. Podemos obter a equação de equilíbrio para o estado {000} a partir da definição de que "a taxa com que o sistema entra no estado n deve ser igual a taxa com que o sistema sai do estado n" (Souza, 2010), da seguinte forma, conforme a equação 1:

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

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

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

Em que:

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

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

$$\boldsymbol{\{111\} \rightarrow \mu P_{\{111\}} = \lambda P_{\{110\}} + \lambda P_{\{101\}} + \lambda P_{\{011\}}} \quad (3)$$

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

$$\boldsymbol{P_{\{000\}} + P_{\{001\}} + P_{\{010\}} + P_{\{100\}} + \cdots + P_{\{111\}} + P_{\{S_4\}} + P_{\{S_5\}} + \cdots = 1} \quad (4)$$

Como dito anteriormente, diversas medidas de desempenho podem ser calculadas a partir das probabilidades do sistema estar em cada estado, obtidas pela solução das equações de equilíbrio do sistema. Essas medidas auxiliam na configuração e análise do sistema sob a hipótese de que o sistema esteja em equilíbrio (LARSON & ODONI, 2007).

## 3 MODELO HIPERCUBO CONSIDERANDO PRIORIDADE NA FILA

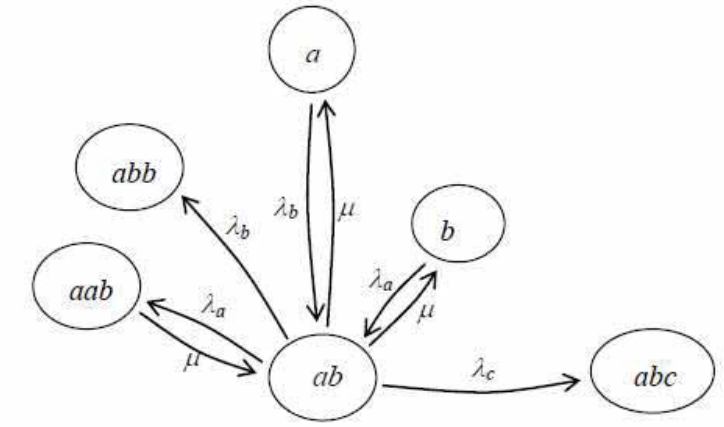
Muitos dos sistemas reais comentados na Seção 1 trabalham considerando classes de usuários em que há uma hierarquia no atendimento. Usuários de classes de maior importância são atendidos antes de usuários de classes de menor importância. Dessa forma, as prioridades são definidas em um determinado número de classes. Para representar esta situação, os átomos geográficos precisam ser divididos em camadas (*layers*) de forma que cada camada represente uma prioridade específica do sistema (Takeda *et. al.*, 2007 e Souza *et. al.*, 2013). Assim, a prioridade envolve a ordem pela qual os chamados são atendidos na fila e o atendimento não se dá pela sequência FCFS, mas sim de acordo com alguma regra como em Souza *et. al.*, (2013, 2015) que considera as classes de acordo com a gravidade envolvida no chamado.

A Figura 2 mostra um exemplo apresentado em Souza *et. al.*, 2013) de como ocorrem as mudanças de um estado de fila onde há prioridade. Note que a classe de chamados *a* tem prioridade em relação à classe de chamados *b*, assim como a classe *b* tem prioridade a classe *c*. Isso, visto que enquanto houve um chamado *a* em fila o chamado *b* não será atendido. Por exemplo, a equação de equilíbrio do estado $\{ab\}$ é dada por:

$$(\lambda+\mu)P_{\{ab\}} = \lambda_a P_{\{b\}} + \lambda_b P_{\{a\}} + \mu P_{\{aab\}} \quad (5)$$

onde $\lambda = \sum_{r\in D} \lambda_r$ e $P_{\{ab\}}$ é a probabilidade em equilíbrio do estado $\{ab\}$.

Figura 2 – Mudanças de estado em fila com prioridade
Fonte: Souza *et. al.* (2015).




Ainda, segundo Iannoni (2005), existem sistemas caracterizados por possuírem políticas de despacho particulares, onde, pelo menos um átomo não é atendido por todos os servidores do sistema. Assim, alguns átomos só podem ser atendidos por determinados servidores. Isso ocorre, em particular, em sistemas onde os tempos de viagem são longos, ou seja, onde a viagem para os átomos se dá através de rodovias, por exemplo. Os sistemas de atendimento emergencial em rodovias estudados não admitiam fila de espera.

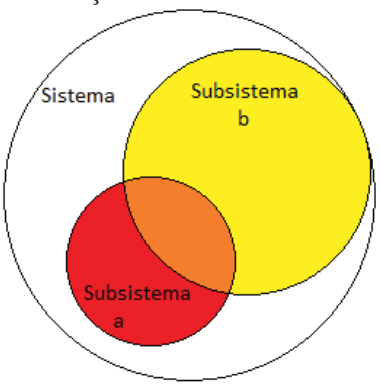
Maiores detalhes sobre modelos que considera prioridade e fila FCFS podem ser encontrados em Takeda et. al. (2007), que utiliza a estratégia de *layering*. Os modelos que consideram prioridade na fila

podem ser encontrados em Souza et. al. (2014, 2015). Além disso, recentemente em Rodrigues (2014) contém detalhes dos modelos que consideram prioridade na fila e backup parcial em que os servidores não podem se deslocar para qualquer átomo.

### 3.1 Modelo hipercubo considerando Backup Parcial e prioridade na fila

Uma maneira de representar esses sistemas é entendendo-os como um conjunto de subsistemas menores, onde as intersecções são os casos onde ambos sistemas operam juntos, enquanto os locais onde não há contato são os locais de ocorrência do *backup* parcial, a Figura 3 ilustra essa ideia. Note que o espaço em laranja representa o ponto em que os sistemas operam conjuntamente, enquanto as áreas em vermelho e amarelo são os pontos em que ocorre o *backup* parcial. Estamos supondo que os servidores atendem em qualquer átomo e o *backup* parcial ocorre apenas no tipo do chamado, alguns servidores estão habilitados a atender apenas a um conjunto limitado dos chamados e não o conjunto todo. Uma representação desse sistema pode ser vista no Anexo 1.

Figura 3 – Ilustração de um sistema com backup parcial




As equações de equilíbrio são construídas supondo-se que o sistema atinja o estado estável (*steady state*). Visto o estado de equilíbrio, o fluxo de entrada em qualquer estado deve ser igual ao fluxo de saída. Outro ponto a ser considerado, em sistemas que possuem *backup* parcial, é que, para o sistema entrar em equilíbrio (*steady state*), é necessário que cada subsistema tenha uma taxa de chegada menor do que a taxa de serviço total dos servidores que podem atender aos chamados daquele subsistema. Em outras palavras, não basta o sistema como um todo ter uma taxa de chegada de chamados menor do que a taxa total de serviço, é preciso analisar as subdivisões do sistema individualmente.

Vale ressaltar que em sistemas com *backup* parcial haverá estados de não saturação (onde não há filas ou perdas com a chegada de um chamado), estados de saturação (onde qualquer novo chamado, ou resultará em fila, ou em perda) e estados de semi-saturação (onde a chegada de chamados de áreas específicas causarão fila ou perda, enquanto a chegada de outros chamados poderá ser atendida de imediato). Com isso em vista, é possível que o sistema apresente estados de fila mesmo quando há servidores disponíveis. Isso pode ocorrer em estados de semi-saturação, quando há a chegada de um chamado em que o servidor livre não pode atendê-lo. Isso resulta em outro problema, na formulação de estados de fila, alguns chamados que se encontram em fila, podem não ser atendidos conforme algum servidor se torna disponível. Assim, para fins de modelagem é necessária a distinção entre as filas.

As medidas de desempenho podem ser calculadas de diferentes maneiras, conforme a existência ou não de fila no sistema. Neste trabalho, apresentar-se-á apenas as medidas para sistemas onde ocorrem filas.

Primeiramente, a carga de trabalho, $\rho_n$, para cada servidor ($n$) pode ser obtida pela soma das probabilidades dos estados em que o dado servidor está ocupado (Larson & Odoni, 2007).

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

$$f_{ij} = f_{ij}^{nq} + f_{ij}^{q} = \frac{\lambda_j}{\lambda} \cdot \sum_{B \epsilon E ij} P_B + \frac{\lambda_j}{\lambda} \cdot \sum_{E} \left( \frac{\sum_{S \epsilon T_E} \lambda_S}{\lambda} \cdot P_E \right) \cdot \frac{\mu_i}{\sum_{A \epsilon M j} \mu_A} \quad (6)$$

Onde $E_{ij}$ é o conjunto de estados em que um chamado do átomo *j* seja atendido prontamente pelo servidor *i*. $M_j$ é o conjunto de servidores habilitados a atenderem aos chamados do átomo *j*. Além disso, $\sum_E \left( \frac{\sum_{S \epsilon T_E} \lambda_S}{\lambda} \cdot P_E \right)$ é a probabilidade de saturação do sistema. Contudo, como apresentado anteriormente, existem estados de semi-saturação, onde apenas os chamados de certos átomos incorrem em fila, ou seja, para certos chamados ele é considerado saturado, mas para outros é considerado não saturado. $T_E$ é o conjunto de átomos cujo chamado ocasionará em fila ou perda, caso esteja no estado *E*. Vale ressaltar que, caso o estado seja de saturação, a probabilidade deste será a mesma que a de seu estado, visto que $\sum_{S \epsilon T_E} \lambda_S$ será igual a λ. Assim sendo, para estados de semi-saturação, apenas uma fração da probabilidade do estado será considerada, a fração relativa à taxa de chegada dos átomos em que resultarão filas em relação à taxa total de chegada. Estados não saturados terão probabilidade 0. Vale ressaltar ainda que os servidores que não estão presentes em $M_j$ terão sua $f_{ij}$ igual a zero em ambos os termos. Além disso, em sistemas que possuem perda, a taxa de entrada de usuários é dada por:

$$\bar{\lambda} = \lambda \left( 1 - \sum_{E} \left( \frac{\sum_{S \epsilon T_E} \lambda_S}{\lambda} \cdot P_E \right) \right) \quad (7)$$

O tempo médio de viagem de um servidor *i* para o subátomo *jr*, é $t_{i,jr}$, dado pela equação (7). $l_{i,jr}$ representa a probabilidade de o servidor *i* estar localizado no subátomo *jr*. O termo $\tau_{jr,pl}$ é o tempo de viagem do subátomo *pl* para o subátomo *jr* (Souza, 2010). Lembrando que, caso o servidor *i* não esteja habilitado a viajar para o subátomo *pl*, não faz sentido calcular seu tempo de viagem para este subátomo.

$$t_{i,jr} = \sum_{p=1}^{N_A} \sum_{l \in D} l_{i,jr} \cdot \tau_{pl,jr} \quad (8)$$

O tempo médio de viagem para chamados em fila, $\bar{T}_Q$, é dado pela equação (8).

$$\bar{T}_Q = \sum_{p} \sum_{l} \sum_{j} \sum_{r} \frac{\lambda_{pl} \lambda_{jr}}{\lambda^2} \tau_{pl,jr} \quad (9)$$

O tempo médio de viagem ao subátomo *jr* é dado pela equação (9). $P_S$ é a probabilidade de saturação apresentada na equação (7) sob o termo $\sum_E \left( \frac{\sum_{S \epsilon T_E} \lambda_S}{\lambda} \cdot P_E \right)$.

$$\bar{T}_{jr} = \frac{\sum_i f_{i,jr}^{nq} t_{i,jr}}{\sum_i f_{i,jr}^{nq}} (1 - P_S) + \sum_{p} \sum_{l} \left( \frac{\lambda_{pl}}{\lambda} \right) \tau_{jr,pl} P_S \quad (10)$$

O tempo médios de viagem do servidor *i*, $\overline{TU}_i$, é obtido pela equação (11).

$$\overline{TU}_i = \frac{\sum_j \sum_r f_{i,jr}^{nq} t_{i,jr} + \left( \bar{T}_Q P_S \right) \frac{\mu_i}{\mu}}{\sum_j \sum_r f_{i,jr}^{nq} + P_S \frac{\mu_i}{\mu}} \quad (11)$$

Para o cálculo dos tempos médios em fila, utiliza-se a fórmula de Little (1961), descrita na equação (12), abaixo. Observe que $L_{qr}$ representa a soma das probabilidades com fila da prioridade *r*, calculados assim como em Souza et. al. (2015). Enquanto que $\sum_E \left( \frac{\sum_{S \epsilon T_E} \lambda_S}{\lambda} \cdot P_E \right)$ corresponde à probabilidade de perda do sistema. $T_L$ é o conjunto de estados (semi-saturados e saturados) onde há possibilidade de ocorrem perdas, lembrando que a probabilidade é proporcional à taxa de chegada dos átomos que ocorrerão perdas, em estados semi-saturados, uma fração dos chamados, enquanto para estados saturados, os chamados em sua totalidade.

$$W_{qr} = \frac{L_{qr}}{\bar{\lambda}} \quad (12)$$

### 3.2 Exemplo Ilustrativo

Considere um sistema de emergência operando em uma região representada por três átomos, utilizando política de despacho de preferência fixa, mostrada na Tabela 2. Observe que para os subátomos "*b*" não foram alocados os servidores 1 e 2, que atendem apenas a chamados com prioridade "*a*". A matriz de preferência de despacho também leva em conta a localização dos servidores, os servidores 1,2 e 3 estão localizados no átomo 1 e o servidor 4 está localizado no átomo 2. Os tempos de viagem entre os subátomos do sistema são dados pela Tabela 3. As taxas de chegadas aos subátomos são: $\lambda_{1a}$ = 0,5, $\lambda_{1b}$ = 1,0, $\lambda_{2a}$ = 0,6, $\lambda_{2b}$ = 1,2, $\lambda_{3a}$ = 0,4 e $\lambda_{3b}$ = 1,3. As taxas de serviço dos servidores são: $\mu_1$ = 1,5, $\mu_2$ = 1,5, $\mu_3$ =2,0 e $\mu_4$ = 2,0

Tabela 2: Matriz de preferência de despacho considerando *backup* parcial

| **Átomos** | **1º** | **2º** | **3º** | **4º** |
|---|---|---|---|---|
| **1a** | 1 | 2 | 3 | 4 |
| **1b** | 3 | 4 | - | - |
| **2a** | 2 | 1 | 3 | 4 |
| **2b** | 4 | 3 | - | - |
| **3a** | 1 | 2 | 3 | 4 |
| **3b** | 3 | 4 | - | - |

Tabela 3: Tempos de viagem entre subátomos

| **Subátomos** | **1a** | **1b** | **2a** | **2b** | **3a** | **3b** |
|---|---|---|---|---|---|---|
| **1a** | 5 | 5 | 10 | 10 | 7 | 7 |
| **1b** | 5 | 5 | 10 | 10 | 7 | 7 |
| **2a** | 10 | 10 | 5 | 5 | 11 | 11 |
| **2b** | 10 | 10 | 5 | 5 | 11 | 11 |
| **3a** | 7 | 7 | 11 | 11 | 5 | 5 |
| **3b** | 7 | 7 | 11 | 11 | 5 | 5 |

Com todas essas informações, pode-se, agora, analisar a capacidade do sistema em atender aos chamados, ou seja, a capacidade de entrar no estado de equilíbrio (*steady state*). Observando a separação do sistema de acordo com as prioridades dos chamados e, além disso, a presença de *backup* parcial que leva em conta essas prioridades, já que os servidores avançados (1 e 2) não atendem chamados básicos (subátomos "*b*"). Isso, caracteriza a presença de dois subsistemas, o subsistema "*a*" e o subsistema "*b*". Então, primeiramente, é preciso totalizar as taxas de chegada de cada subsistema. Para o subsistema "*a*", tem-se a taxa total de chegada $\lambda_a$, já para o subsistema "*b*", tem-se a taxa total de $\lambda_b$. Por outro lado, também precisa-se somar as taxas de serviço disponíveis para cada subsistema. Para o subsistema "*a*", já que todos servidores estão disponíveis para atendê-lo, a taxa total de serviço será a soma da taxa de todos servidores, totalizando um $\mu_{Sp1}$. Enquanto, para o subsistema "*b*", apenas os servidores 3 e 4 estão habilitados a atender às suas ocorrências, totalizando um $\mu_{Sp2}$. A Tabela 4 mostra o resultado da verificação de capacidade.

Tabela 4: Verificação de capacidade de cada subsistema no exemplo 2.

| **Verificação** | | | | |
|---|---|---|---|---|
| $\lambda a$ = | 1,5 | < | $\mu Sp1$ = | 7,0 |
| $\lambda b$ = | 3,5 | < | $\mu Sp2$ = | 4,0 |

Essa verificação mostra que ambos subsistemas têm capacidade de entrarem em estado de equilíbrio, já que possuem maior taxa de serviço do que taxa de chegada de chamados. É importante ter em vista que

caso $\lambda_a$ ou $\lambda_b$ fossem maiores do que $\mu_{Sp1}$ ou $\mu_{Sp2}$, respectivamente, o sistema não teria condições de estar em equilíbrio, já que chegariam mais chamados do que o sistema os atenderia, não fazendo sentido em construir e resolver as equações de equilíbrio, ou seja, encontrar as probabilidades de estado. Essa observação mostra que a análise transitória, assim como realizada em Souza *et. al.* (2014), também pode ser importante em casos em que há classes de usuários e algum subsistema não entra em equilíbrio.

O próximo passo, então, é construir as equações de equilíbrio do sistema. Primeiramente precisa-se realizar a construção dos estados do sistema. Observe que o estado 0011*bb* é um estado de semi-saturação uma vez que com a chegada de um chamado dos subátomos "a" esta será prontamente atendida pelos servidores avançados (1 ou 2), já que estes estão disponíveis, levando aos estados 0111*bb* ou 1011*bb*. Contudo, a chegada de um chamado dos subátomos "*b*" não será atendido de imediato, já que os servidores avançados não podem atendê-los e os servidores básicos (3 e 4) já estão ocupados, levando assim o sistema para o estado 0011*bbb* (um chamado "*b*" a mais na fila). Alguns exemplos das equações de equilíbrio desse sistema estão apresentadas na Tabela 5. Note que no estado 1000 tem-se a situação de não saturação, visto que qualquer chamado pode ser atendido de imediato, já que se possui um servidor avançado livre e, especialmente, os dois servidores básicos também (atendem aos dois subsistemas). Por outro lado, o estado 0011*bb* é um estado de semi-saturação, como mostrado anteriormente. Por fim, o estado 1111*ab* é um estado de saturação, já que um chamado de qualquer subsistema leva ao aumento da fila, visto que todos servidores estão ocupados.

Tabela 5: Algumas equações de equilíbrio do sistema

| Estados | Saída | | Entrada |
|---|---|---|---|
| 1000 | $(\lambda+\mu_1)*P_{1000}$ | = | $\mu_2*P_{1100} + \mu_3*P_{1010} + \mu_4*P1001 + (\lambda_{1a}+\lambda_{3a})*P_{0000}$ |
| 0011bb | $(\lambda+\mu_3+\mu_4)*P_{0011bb}$ | = | $\mu_1*P_{1011bb} + \mu_2*P_{0111bb} + (\mu_3+\mu_4)*P_{0011bbb} + \lambda_b*P_{0011b}$ |
| 1111ab | $(\lambda+\mu)*P_{1111ab}$ | = | $\mu*P_{1111aab} + \lambda_a*P_{1111b} + \lambda_b*P_{1111a}$ |

O sistema pode ser resolvido a partir do método de Gauss-Jordan. Uma vez resolvido e as probabilidades obtidas, pode-se calcular as medidas de desempenho do sistema. Primeiramente, serão calculadas as cargas de trabalho dos servidores, a Tabela 6 mostra as cargas de trabalho dos servidores.

Tabela 6: Carga de trabalho dos servidores

| Ambulância | Carga de Trabalho |
|---|---|
| 1 | 0,4355 |
| 2 | 0,3965 |
| 3 | 0,8021 |
| 4 | 0,7773 |
| **Sistema** | **0,6028** |

A Tabela 7 mostra a frequência de despacho, calculadas a partir da equação (6), note que tanto a probabilidade quanto a frequência de envio dos servidores avançados para os subátomos "*b*" são sempre iguais a 0.

Tabela 7: Frequência de despacho dos servidores

| $f_{i,jr}$ | 1ª | 1b | 2ª | 2b | 3a | 3b |
|---|---|---|---|---|---|---|
| **1** | 0,0678 | 0,0000 | 0,0347 | 0,0000 | 0,0543 | 0,0000 |
| **2** | 0,0328 | 0,0000 | 0,0861 | 0,0000 | 0,0262 | 0,0000 |
| **3** | 0,0185 | 0,0927 | 0,0222 | 0,0852 | 0,0148 | 0,1206 |
| **4** | 0,0173 | 0,0760 | 0,0208 | 0,1173 | 0,0139 | 0,0988 |

A Tabela 8 mostra os números médios de usuários no sistema, a taxa de entrada e os tempos médios de espera em fila, calculados a partir da equação (12), para cada classe de usuários e no sistema.

Tabela 8: Números médios de usuários no sistema, a taxa de entrada e os tempos médios de espera em fila.

| **Subsistema** | $L_Q$ | $\bar{\lambda}$ | $W_{qr}$ |
|---|---|---|---|
| ***a*** | 0,0266 | 1,3220 | 0,0202 |
| ***b*** | 0,1165 | 3,0847 | 0,0378 |
| **Geral** | 0,1432 | 4,4068 | 0,0325 |

Os tempos médios de viagem dos servidores aos subátomos foram calculados a partir da equação (8). Seus resultados estão na Tabela 9.

Tabela 9: Tempos de viagem dos servidores aos subátomos.

| $t_{i,jr}$ | **Subátomos** | | | | | |
|---|---|---|---|---|---|---|
| **Servidores** | **1a** | **1b** | **2a** | **2b** | **3a** | **3b** |
| **1** | 5 | - | 10 | - | 7 | - |
| **2** | 5 | - | 10 | - | 7 | - |
| **3** | 5 | 5 | 10 | 10 | 7 | 7 |
| **4** | 10 | 10 | 5 | 5 | 11 | 11 |

Seguindo com o tempo médio de viagem para chamados em fila, tem-se $\bar{T}_Q = 7{,}96$, obtido através da equação (9). Os tempos médios de viagem aos subátomos, obtidos pela equação (10), e os tempos médios de viagem dos servidores, calculados pela equação (11), estão presentes na Tabela 10.

Tabela 10: Tempos de viagem aos subátomos

| **Subátomo** | $\bar{T}_{jr}$ | **Servidores** | $\overline{TU}_i$ |
|---|---|---|---|
| **1a** | 6,38 | **1** | 7,25 |
| **1b** | 7,17 | **2** | 8,26 |
| **2a** | 9,16 | **3** | 7,49 |
| **2b** | 7,55 | **4** | 7,88 |
| **3a** | 7,45 | | |
| **3b** | 8,09 | | |

A extensão considerada neste trabalho do modelo hipercubo para considerar prioridade na fila e *backup* parcial nos tipos de chamados é um caso importante a ser considerado em sistemas como os SAMU´s, em que é permitida fila de espera existem servidores dedicados a um tipo de chamado e servidores que atendem todos os tipos de chamados. Além disso, todos os servidores podem viajar para qualquer átomo na região.

## 4 CONCLUSÃO

Dada a relevância do modelo hipercubo na análise de Sistemas de Atendimento Emergenciais e observando que estes sistemas possuem diferentes características que devem ser incorporadas no modelo a fim de se obter resultados com mais acurácia para serem comparados aos sistemas reais, o objetivo desse trabalho é estender o modelo hipercubo para considerar prioridade na fila e backup parcial quanto às classes dos chamados e não há restrição física. Ou seja, os servidores podem viajar a qualquer átomo geográfico para atender um chamado. Foram realizadas modificações nas medidas de desempenho para atender às hipóteses da extensão proposta.

Além disso, uma importante observação deste trabalho é em relação a capacidade de atendimento do sistema, mesmo que a taxa de entrada total seja menor que a taxa de serviço total é necessária a

verificação mostra que ambos subsistemas (classes a e b) têm capacidade de entrarem em estado de equilíbrio. Essa observação mostrou a importância da análise transitória, assim como realizada em Souza et. al. (2014), deve ser investigada nesses tipos de sistemas uma vez que em casos em que há classes de usuários e algum subsistema pode não entrar em equilíbrio. Dessa forma, uma perspectiva futura a ser considerada é a análise transiente em sistemas reais a fim de verificar o impacto da abordagem clássica do modelo hipercubo e da abordagem transiente quando somente uma parte do sistema entra em equilíbrio.
Outra perspectiva é a aplicação dessa abordagem em SAE´s, como os SAMU´s, a fim de verificar a acurácia do modelo.

**REFERÊNCIAS BIBLIOGRÁFICAS**

**ALANIS R., INGOLFSSON A., KOLFAL, B.** A Markov chain model for an EMS system with repositioning. *Production and Operations Management*, 22(1), 216-231, 2013.

**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. *TIMS Studies in the Management Science*, 22, Elsevier, 121-153, 1986.

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

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

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

**CHIYOSHI F., IANNONI A. P., MORABITO R.** A tutorial on hypercube queuing models and some practical applications in emergency service systems. *Pesquisa Operacional*, 31( 2), 271-299, 2011.

**CHIYOSHI F., MORABITO R.,** "A Tabu search algorithm for solving the extended maximal availability location problem", *International Transactions in Operational Research*, 18, 6, 663-678, 2011.

**CHIYOSHI, F., IANNONI, A. P. E MORABITO, R.,** "A tutorial on hypercube queueing models and some practical applications in emergency service systems", *Pesquisa Operacional*, 31, 2, 271-299, 2011.

**COSTA D.M.** Uma metodologia iterativa para determinação de zonas de atendimento de serviços emergenciais. Universidade Federal de Santa Catarina. *Tese* (doutorado em Engenharia de Produção) - Departamento de Engenharia de Produção, 2004.

**FIGUEIREDO A. P. S., LORENA L. A. N.** Localização de ambulâncias: uma aplicação para a cidade de São José dos Campos. *Anais* XII Simpósio Brasileiro de Sensoriamento Remoto, Goiânia, Brasil, INPE, p. 1965-1972, 2005.

**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*, 15, 525-549, 2008.

**GONÇALVES M. B., NOVAES A. G., SCHMITZ R.** Um modelo de otimização para localizar unidades de serviço emergenciais em rodovias. In: Congresso de Pesquisa e Ensino em Transportes 9, São Carlos, SP, 1995. *Anais*. São Carlos 3, p.962-972, 1995.

**IANNONI A. P., CHIYOSHI F., MORABITO, R**., "A spatially distributed queuing model considering dispatching policies with server reservation", *Transportation Research*, E 75, 49-66, 2015.

**IANNONI A. P., MORABITO R., SAYDAM C.,** "A hypercube queueing model embedded into a genetic algorithm for ambulance deployment on highways", *Annals of Operations Research* 157 (1), 207–224, 2008.

**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*, 195, 528-542, 2009.

**IANNONI, A. P.** Otimização da configuração e operação de sistemas médico emergenciais em rodovias utilizando o modelo hipercubo. Universidade Federal de São Carlos. *Tese* (doutorado em Engenharia de Produção) - Departamento de Engenharia de Produção, 2005.

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

**LARSON R.C., ODONI A.R.** *Urban Operations Research*. 2 ed. Dynamic Ideas, Belmont, Massachusetts, 2007.

**MENDONÇA F., MORABITO R.,** "Aplicação do modelo hipercubo para análise de um sistema médico-emergencial em rodovia", *Gestão & Produção* 7(1), 73-91, 2000.

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

**RODRIGUES L. F.** Análise dos serviços emergenciais de manutenção agrícola e borracharia na agroindústria canavieira utilizando teoria de filas. Universidade Federal de São Carlos. *Tese* (doutorado em Engenharia de Produção) - Departamento de Engenharia de Produção, 2014.

**SIMPSON N.C., HANCOCK P.G.** Fifty years of operational research and emergency response. Journal of the Operational Research Society 60, p. 126-139, 2009.

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

**SOUZA R. M. , MORABITO R., CHIYOSHI F. Y., IANONNI A. P.** Análise da configuração de SAMU utilizando múltiplas alternativas de localização de ambulâncias. *Gestão & Produção*, 20(2), 287-302, 2013.

**SOUZA R. M. , MORABITO R., CHIYOSHI F. Y., IANONNI A. P.** Abordagem dinâmica nos modelos M(t)/M(t)/m(t)/C(t) e hipercubo. *Anais* SBPO, 2014.

**SOUZA R. M. , MORABITO R., CHIYOSHI F. Y., IANONNI A. P.** Incorporating priorities for waiting customers in the Hypercube Queuing Model, with application to an emergency medical service system in Brazil. *European Journal of Operational Research*, v. 242, p. 274-285, 2015.

**SOUZA, R. M., MORABITO, R., CHIYOSHI, F. E IANNONI, A. P.,** "Extensão do modelo hipercubo para análise de sistemas de atendimento médico emergencial com prioridade na fila", *Produção*, 24, 1, 1-12, 2014.

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

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