O USO DO MODELO HIPERCUBO NA SOLUÇÃO DE PROBLEMAS DE LOCALIZAÇÃO PROBABILÍSTICOS

Fernando Chiyoshi
Roberto D. Galvão
Programa de Engenharia de Produção – COPPE/UFRJ
Caixa Postal 68507 – 21945-970 – Rio de Janeiro – RJ

Reinaldo Morabito
Departamento de Engenharia de Produção – UFSCar
Caixa Postal 676 – 13565-905 – São Carlos – SP

Resumo

O modelo hipercubo é revisitado tendo em vista sua utilização em métodos de solução para problemas de localização probabilísticos. Este uso do modelo é de bastante relevância em situações em que a aleatoriedade na disponibilidade dos servidores é um fator importante a ser considerado; em algumas circunstâncias esta aleatoriedade só pode ser representada pela modelagem de filas espacialmente distribuídas. O modelo é apresentado com o auxílio de um exemplo ilustrativo, para o qual são derivadas as equações de equilíbrio; medidas de desempenho do modelo são também definidas. Isto é seguido pela descrição de um método exato e de outro aproximado para o cálculo destas medidas. Diversos modelos de localização probabilísticos são então estudados, o que é seguido pela análise de métodos de solução disponíveis para esses modelos, com ênfase especial nos métodos que incluem o uso do modelo hipercubo. Embora atualmente de uso incipiente em problemas de localização probabilísticos, o modelo tem grande potencial nesse contexto, por exemplo se integrado a metaheurísticas tais como simulated annealing *e busca tabu.*

Palavras-chave: localização, modelo hipercubo, modelos probabilísticos, filas.

1. Introdução

O interesse em problemas de localização tem crescido bastante a partir da década de 60, quando foram definidos os primeiros modelos matemáticos nessa área; a literatura correspondente é atualmente bastante vasta. Os modelos de localização normativos, assim chamados

porque caracterizam-se pela otimização de uma *norma* ou *medida de utilidade*, sujeita às restrições operacionais relevantes, correspondem a problemas que podem ser formulados e resolvidos com base em técnicas de otimização de modelos matemáticos.

Problemas de localização no setor público podem ser classificados em duas categorias: localização de serviços não emergenciais e localização de serviços de emergência. Na primeira categoria estão incluídos a localização de escolas, agências de correio, edifícios públicos, de alguns serviços de saúde pública e mesmo de serviços relacionados ao meio-ambiente, como suprimento de água e facilidades para o depósito de lixo. A categoria de serviços de emergência inclui por exemplo a localização de hospitais, de serviços de atendimento de emergência por ambulâncias e de estações do corpo de bombeiros. As medidas de utilidade a serem otimizadas são diferentes para as duas categorias. No caso de serviços de emergência uma medida de utilidade bastante usada é a máxima distância a ser percorrida (chamada de *distância crítica*) entre qualquer usuário do sistema e a facilidade (ou servidor) mais próxima.

Os primeiros modelos a serem desenvolvidos para a localização de serviços de emergência foram modelos determinísticos. Uma desvantagem dos modelos determinísticos é que eles partem da hipótese que os servidores estão disponíveis quando solicitados, o que nem sempre é razoável em aplicações práticas. O congestionamento em serviços de atendimento de emergência, que pode causar a indisponibilidade de um servidor a menos da distância crítica quando solicitado, motivou o desenvolvimento dos modelos de localização probabilísticos. Na modelagem probabilística de serviços de emergência, algumas hipóteses simplificadoras permitem o uso da programação matemática. Situações em que hipóteses simplificadoras não são aplicáveis conduzem, no entanto, ao tratamento desses problemas através de programação estocástica.

O modelo hipercubo, proposto por LARSON (1974) e estudado por diversos autores (SWERSEY, 1994), é uma importante ferramenta para o planejamento de sistemas de serviços, especialmente os sistemas urbanos em que servidores se deslocam para fornecer algum tipo de serviço para clientes (*server-to-customer service*). O modelo é adequado para analisar sistemas coordenados ou centralizados, onde o usuário que deseja algum serviço (p.e., atendimento médico ou policial, entrega de algum item como uma pizza ou uma peça de computador) telefona para a central de atendimento do sistema. O administrador do sistema então despacha um servidor de alguma facilidade próxima do local da chamada para atender o cliente. Caso nenhum servidor esteja disponível, a solicitação entra numa fila de espera para ser atendida assim que algum servidor ficar disponível.

O modelo aborda complexidades geográficas e temporais da região, com base em resultados de teoria de filas espacialmente distribuídas e aproximações a partir de análise Markoviana. Basicamente, a idéia é expandir a descrição do espaço de estados de um sistema de fila com múltiplos servidores, para poder representar cada servidor individualmente e incorporar políticas de despacho mais complexas. A solução do modelo envolve resolver um sistema de equações lineares que fornece as probabilidades de equilíbrio dos possíveis estados do sistema (*probabilidades de estado*). Estas probabilidades permitem estimar diversas medidas de desempenho interessantes para o gerenciamento do sistema, tais como cargas de trabalho (*workloads*) dos servidores, tempo médio de resposta do sistema ou de cada servidor, freqüência de atendimento de cada servidor a cada região, entre outras. Apesar de ser aplicável a estudos de cenários, o modelo hipercubo é um modelo descritivo que, se aplicado de forma isolada, não assegura a solução direta de problemas de localização.

Nosso interesse no presente trabalho é revisitar o modelo hipercubo e enfatizar a importância de sua integração à modelagem de problemas de

Tabela 1 – Lista de preferências de despacho.

Átomo	Servidores Preferenciais 1º.	2º.	3º.
1	1	2	3
2	2	3	1
3	3	1	2

localização nos quais a aleatoriedade na disponibilidade dos servidores é um fator importante a ser considerado. Esta integração possibilita a modelagem de situações mais próximas das encontradas no mundo real, através do tratamento matemático de filas espacialmente distribuídas. O trabalho tem caráter tutorial quanto ao modelo hipercubo, e caráter de revisão bibliográfica no que se refere aos modelos de localização probabilísticos apresentados.

O modelo hipercubo é tratado em detalhe na seção 2 com o auxílio de um exemplo ilustrativo, para o qual são derivadas as equações de equilíbrio; medidas de desempenho do modelo são também discutidas nesta seção. Isto é seguido pela descrição, na seção 3, de métodos exatos e aproximados para o cálculo das medidas de desempenho. Em particular, discute-se aspectos práticos da implementação de métodos iterativos, o que foi pouco reportado na literatura, como a importância dos valores iniciais das probabilidades de equilíbrio para a convergência dos métodos. Também apresenta-se uma experiência computacional da aplicação dos métodos em exemplos aleatórios. Diversos modelos de localização probabilísticos são então estudados na seção 4, o que é seguido (seção 5) da análise de métodos de solução disponíveis para esses modelos, com ênfase especial nos métodos que incluem o uso do modelo hipercubo. As conclusões encerram o artigo.

2. O Modelo Hipercubo

O modelo hipercubo baseia-se na divisão da região em um conjunto de áreas de demanda ou *átomos geográficos*, cada átomo considerado como uma fonte pontual independente solicitadora de serviços ao longo do tempo (similarmente aos nós ou vértices como pontos de demanda de uma rede de transportes). Desta maneira, o modelo considera os chamados distribuídos tanto espacialmente quanto temporalmente na região. O atendimento aos chamados de cada átomo é realizado por *servidores* (ambulâncias, veículos de entrega, bombeiros) distribuídos na região e que, quando disponíveis, podem estar fixos em alguns pontos (bases), ou movimentando-se em sub-regiões (neste caso seus movimentos devem ser conhecidos ao menos probabilisticamente). Para um servidor, sua *área de cobertura primária* é o conjunto dos átomos que este servidor atende prioritariamente (i.e., é o primeiro a ser chamado; o modelo também trata situações em que podemos ter átomos com mais de um servidor preferencial). No caso de o servidor preferencial estar ocupado, outros servidores são chamados para atender a solicitação.

O nome hipercubo deriva da descrição da disponibilidade dos servidores por meio do espaço de estados. Cada servidor pode estar livre (0) ou ocupado (1) em um certo instante. Um estado particular do sistema é dado pela lista dos servidores que estão livres e ocupados. Por exemplo, o estado 110 corresponde a um sistema com três servidores, com o servidor 1 livre e os servidores 2 e 3 ocupados (note que 110 descreve o estado dos servidores da direita para a esquerda). Desta maneira, para três servidores o espaço de estados do sistema é dado pelos vértices de um *cubo*. No caso de termos mais de

três servidores, temos um *hipercubo*. O modelo trata tanto sistemas em que os chamados não atendidos aguardam em fila, quanto sistemas que não admitem a formação de filas.

Alguns exemplos de aplicação do modelo hipercubo nos EUA são a localização de ambulâncias em Boston, MA (BRANDEAU & LARSON, 1986), o patrulhamento policial em Orlando, FL (SACKS & GRIEF, 1994), e o programa de visitas do serviço social (LARSON & ODONI, 1981). No Brasil, alguns exemplos são o atendimento a interrupções na distribuição de energia elétrica em Santa Catarina (ALBINO, 1994), a localização de ambulâncias em um trecho da BR-111 (GONÇALVES *et al.*, 1994, 1995), o balanceamento da carga de trabalho de ambulâncias no sistema "Anjos do Asfalto" da Via Dutra (MENDONÇA, 1999, e MENDONÇA & MORABITO, 2000), e a configuração do Serviço de Atendimento Médico de Urgência (SAMU) da prefeitura de Campinas, SP (TAKEDA, 2000). Extensões do modelo hipercubo foram consideradas em HALPERN (1977) e BURWELL *et al.* (1993). Outras referências podem ser encontradas em SWERSEY (1994). Para a aplicação do modelo, há nove hipóteses básicas que devem ser satisfeitas:

1. *Átomos geográficos*: a região analisada é dividida em N_A átomos geográficos. Os átomos podem corresponder a trechos de uma rodovia, áreas de uma cidade ou de uma região. Em geral cada átomo é modelado com um simples ponto, localizado no centro do átomo real, como em uma rede de transportes.
2. *Processos de chegada conforme processos de Poisson independentes*: admite-se que os chamados de cada átomo j são gerados conforme um processo de *Poisson*, independente dos outros átomos, com taxa média λ_j ($j = 1,2,...,N_A$). Embora aparentemente muito restritiva, esta condição é comumente satisfeita em diversos sistemas reais.
3. *Tempos de deslocamento do servidor*: os tempos médios de viagem τ_{ij} entre o átomo i e o átomo j ($i,j = 1,2,...,N_A$) deverão ser conhecidos ou estimados pelos conceitos de probabilidade geométrica.
4. *Servidores*: existem N servidores espacialmente distribuídos ao longo da região, que podem se deslocar e atender qualquer um dos átomos. Em certos casos essa hipótese pode ser facilmente relaxada para representar políticas de despacho particulares, por exemplo se apenas os dois servidores mais próximos do local do chamado são considerados para atender a solicitação (veja MENDONÇA & MORABITO, 2000).
5. *Localização dos servidores*: cada servidor, quando disponível, pode ficar fixo em um átomo, ou se mover (p.e., em patrulhamento) dentro de uma área determinada (neste caso sua localização deve ser conhecida, ao menos probabilisticamente).
6. *Despacho de um servidor*: apenas um servidor é despachado para atender cada serviço solicitado. O modelo não representa adequadamente situações onde mais de um servidor é despachado para a mesma chamada (p.e., no caso de um grande incêndio), embora em muitas situações reais o conjunto de servidores despachados possa ser visto como um único servidor. Se não houver servidores disponíveis, poderá haver formação de filas (no caso de sistemas que permitem filas), ou perda do chamado (no caso de sistemas que não permitem filas), que então é transferido para outro sistema de atendimento.
7. *Política de despacho dos servidores*: há uma lista fixa de preferências de despacho para cada átomo. Se o primeiro servidor desta lista estiver disponível, ele é despachado para atender o chamado do átomo, caso contrário o próximo servidor disponível da lista (i.e., o *backup*) é despachado. A lista de preferências é fixada *a priori* e permanece inalterada durante a operação do sistema. O modelo pode ser facilmente adaptado para casos em que um átomo tem mais de um servidor preferencial.
8. *Tempo de serviço*: o tempo de serviço de um chamado inclui o tempo de preparo do

servidor (*setup time*), tempo de viagem do servidor até o local onde houve o chamado, o tempo de realização do serviço propriamente dito (em cena), e o tempo de retorno à base. Os servidores têm taxas médias de serviço μ_n (n = 1,..,N), que podem ser diferentes entre os servidores. No caso do sistema permitir formação de filas, o modelo funciona melhor à medida que os tempos médios de serviço se aproximam dos respectivos desvios padrões, isto é, à medida que o processo de serviço tende a ser exponencialmente distribuído. Entretanto, segundo LARSON & ODONI (1981), desvios razoáveis desta hipótese não alteram sensivelmente a precisão do modelo. Se o sistema não permitir filas, esta hipótese é ainda menos necessária, dado que sistemas *M/M/N/N* e *M/G/N/N* têm a mesma distribuição de equilíbrio (GROSS & HARRIS, 1974).

9. *Dependência do tempo de serviço em relação ao tempo de viagem*: variações no tempo de serviço devidas às variações no tempo de viagem são assumidas como sendo de segunda ordem, quando comparadas com as variações dos tempos em cena e/ou *setup* (isso não significa que o tempo de deslocamento deva ser ignorado ao se computar o tempo de serviço). Esta hipótese, que limita a aplicabilidade do modelo, é mais comumente verificada em sistemas urbanos do que em sistemas rurais.

Embora muitos serviços do tipo *server-to-customer* operem com prioridades nos tipos de chamados que recebem, é comum não ser necessário considerar estas prioridades diretamente no modelo para fins de planejamento. Por exemplo, sejam chamadas que necessitam uma unidade de UTI e chamadas que podem ser atendidas por uma unidade comum no sistema SAMU em Campinas (TAKEDA, 2000). Para modelar esta situação, cada átomo é dividido em dois átomos, um gerando chamadas do primeiro tipo e o outro gerando chamadas do segundo, com os átomos com chamadas do primeiro tipo tendo como servidores preferenciais as unidades de UTI. Esta maneira de adaptar o modelo para tipos de chamadas e tipos de unidades permite computar separadamente medidas de desempenho para cada tipo de servidor. Políticas de atendimento com prioridades mais complicadas, no entanto, não podem ser tratadas pelo modelo: por exemplo, interromper um serviço em andamento de uma chamada de baixa prioridade para ir atender uma chamada de maior prioridade. Neste caso, o uso de um modelo de simulação talvez seja mais recomendado.

Na prática, nenhum sistema real adere exatamente a todas as nove hipóteses acima. A decisão de aplicar ou não o modelo hipercubo deve levar em conta o quanto o sistema real não se ajusta à rigidez do modelo, contra as limitações ou complicações do uso de modelos alternativos. A descrição detalhada do modelo está além dos objetivos deste artigo; para mais detalhes, o leitor pode consultar, por exemplo, LARSON (1974) e LARSON & ODONI (1981).

2.1 Exemplo Ilustrativo

A título de ilustração apresentamos a seguir um exemplo simples de uma cidade particionada em N_A = 3 átomos, com demandas distribuídas segundo processos de Poisson independentes, com taxas médias de chamadas λ_1, λ_2 e λ_3, respectivamente (conforme as hipóteses básicas 1 e 2 acima). O sistema dispõe de N = 3 servidores que atendem qualquer um dos átomos com tempos de serviço exponencialmente distribuídos, com taxas médias μ_1, μ_2 e μ_3, respectivamente (conforme as hipóteses 4 e 8 acima. Também admitimos que os tempos de viagem satisfazem a hipótese 9). O fato de termos escolhido um exemplo com apenas 3 servidores é para podermos representar seus possíveis estados em um espaço tridimensional. E o fato de termos apenas 3 átomos no exemplo não elimina a generalidade da estrutura do sistema de equações para sistemas com 3 servidores (ver seção 2.2).

Cada chamada é atendida por apenas um servidor e, quando todos os servidores estão ocupados, o chamado entra em uma fila (de capacidade infinita), à espera do primeiro servidor desocupado (hipótese 6). Há uma lista fixa de preferências de despacho, onde cada átomo tem um único servidor preferencial, conforme a Tabela 1 (hipótese 7). Note, por exemplo, que o átomo 2 tem o servidor 2 como preferencial, seguidos dos servidores 3 e 1 (primeiro e segundo *backups*). A disciplina de atendimento dos chamados em fila é FCFS (*First-Come, First-Served*). Os servidores 1, 2 e 3, quando disponíveis, permanecem no centro dos átomos 1, 2 e 3, respectivamente (hipótese 5) – neste exemplo, os átomos 1, 2 e 3 coincidem com as áreas de cobertura primária dos servidores 1, 2 e 3.

Os possíveis estados em que o sistema pode ser encontrado em um dado instante de tempo são: {000}, {001}, {010}, {100}, {011}, {110}, {101}, {111}, {S_4}, {S_5}, {S_6}, ..., onde S_i, $i \geq 4$, corresponde ao estado em que i usuários estão no sistema (isto é, os três servidores estão ocupados e $(i-3)$ chamadas estão aguardando serviço em fila). Podemos representar o diagrama de estados deste exemplo como um *cubo*, com os vértices correspondendo aos estados {000}, {001}, {010}, {100}, {011}, {110}, {101} e {111}, e com uma *cauda* infinita a partir do estado {111}, composta dos estados {S_4}, {S_5}, {S_6}, ...

2.2 Construção das Equações de Equilíbrio

Para a solução do modelo, constrói-se uma equação de equilíbrio para cada estado do sistema, igualando-se a taxa de transição desse estado para outros estados e a taxa de transição de outros estados para o estado considerado. Iniciando com o estado {000} em que todos os servidores estão livres, tem-se que o sistema passa do estado {000} para o estado {001} quando ocorre uma chamada originada do átomo 1. A taxa de ocorrência deste evento é λ_1. Quando a chamada se origina no átomo 2, a transição se dará para o estado {010}, sendo λ_2 a taxa associada a essa transição. Da mesma forma, quando a chamada se origina no átomo 3, a transição se dará para o estado {100} com taxa λ_3. Em conseqüência, a taxa total de transição do estado {000} para outros estados será de $\lambda = \lambda_1 + \lambda_2 + \lambda_3$. Na realidade, independentemente do estado atual do sistema, qualquer chegada provoca a transição do sistema para fora deste estado.

Considerando-se as transições no sentido inverso, tem-se, em primeiro lugar, que o estado {000} pode ser alcançado a partir do estado {001} quando ocorre a conclusão do atendimento que o servidor 1 está realizando, evento que ocorre com taxa μ_1. O estado {000} pode também ser alcançado a partir do estado {010} pela conclusão do atendimento do servidor 2 (taxa μ_2) ou do estado {100} pela conclusão do atendimento do servidor 3 (taxa μ_3).

A equação de equilíbrio em torno do estado será:

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

onde p_B é a probabilidade de equilíbrio do estado B (p.e., p_{000} é a probabilidade do estado B={000}). Convém observar que, dado que o modelo admite que apenas um servidor é alocado para atender cada chamada (hipótese 6), e que é virtualmente impossível a chegada de duas chamadas simultâneas (hipótese 2), nunca teremos uma transição do estado {000} para, digamos, o estado {011} (uma vez que esta assumiria que dois servidores livres ficam ocupados no mesmo instante). Similarmente, dado que o modelo admite que os servidores operam independentemente, e que é virtualmente impossível que dois servidores ocupados terminem seus serviços simultaneamente (hipóteses 6 e 8), nunca teremos uma transição do estado {011} para o estado {000}.

Quando o sistema se encontra no estado {001}, além do fato de que qualquer chegada levará o sistema para fora deste estado, é possível a transição para o estado {000} via

conclusão do atendimento do servidor 1, de modo que a taxa total de transição do estado {001} para outros estados será de $\lambda+\mu_1$. As transições de outros estados para o estado {001} são: (i) do estado {000} pela chegada de uma chamada do átomo 1, (ii) do estado {011} pela conclusão do atendimento do servidor 2 e (iii) do estado {101} pela conclusão do atendimento do servidor 3. Essas possíveis transições levam à equação:

$$\{001\} \quad (\lambda+\mu_1)p_{001} = \lambda_1 p_{000} + \mu_2 p_{011} + \mu_3 p_{101} \qquad (2)$$

As equações relativas aos estados {010} e {100} têm derivações e estruturas análogas:

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

$$\{100\} \quad (\lambda+\mu_3)p_{100} = \lambda_3 p_{000} + \mu_1 p_{101} + \mu_2 p_{110} \qquad (4)$$

Quando o sistema se encontra no estado {011}, além do fato de que qualquer chegada levará o sistema para outros estados, é também possível a transição para o estado {001}, via conclusão do atendimento do servidor 2, e a transição para o estado {010}, via conclusão do atendimento do servidor 1, de modo que a taxa total de transição do estado {011} para outros estados será de $\lambda+\mu_1+\mu_2$. Para as transições de outros estados para o estado {011}, tem-se em primeiro lugar a transição a partir do estado {001}. A chegada de uma chamada do átomo 2 provocará tal transição. Além disso, uma chamada originada no átomo 1 também provocará a mesma transição, de vez que o servidor 1, que é o servidor preferencial do átomo 1, está ocupado, e seu primeiro "reserva" (*backup*) é o servidor 2. Outras possíveis transições para o estado {011} são: a partir do estado {010}, pela chegada de uma chamada originada no átomo 1, e a partir do estado {111}, pela conclusão do atendimento do servidor 3. Tudo considerado, tem-se a equação:

$$\{011\} \quad (\lambda+\mu_2+\mu_1)p_{011} = (\lambda_1+\lambda_2)p_{001} + \lambda_1 p_{010} + \\ + \mu_3 p_{111} \qquad (5)$$

De forma análoga, tem-se:

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

$$\{110\} \quad (\lambda+\mu_2+\mu_3)p_{110} = (\lambda_3+\lambda_2)p_{010} + \lambda_2 p_{100} + \\ + \mu_1 p_{111} \qquad (7)$$

Chega-se finalmente ao estado {111} em que todos os servidores estão ocupados e não há nenhuma chamada esperando atendimento. Neste estado, em primeiro lugar, qualquer chegada ou conclusão de atendimento de qualquer servidor provocará a transição do sistema para fora deste estado. Em segundo lugar, o estado {111} pode ser alcançado a partir dos estados com dois servidores ocupados através de uma chamada, qualquer que seja sua origem. Nesta situação, o único servidor livre será sempre despachado: como servidor preferencial, como primeiro *backup* ou como segundo *backup*. E por fim, tem-se que o estado {111} pode ser alcançado a partir do estado S_4 (em que estão presentes 4 usuários no sistema, três recebendo atendimento e um na fila de espera) via conclusão do atendimento de qualquer servidor. Fazendo Prob$\{S_4\} = p_4$, tem-se a equação:

$$\{111\} \quad (\lambda+\mu)p_{111} = \lambda p_{011} + \lambda p_{101} + \lambda p_{110} + \mu p_4$$

onde $\mu = \mu_1 + \mu_2 + \mu_3$.

Poderíamos prosseguir com as equações de equilíbrio para os estados S_4, S_5, S_6, ..., o que resultaria em um sistema infinito de equações. Ao invés disso, observamos que a condição de equilíbrio do sistema requer que as taxas de transição entre os estados {111} e $\{S_4\}$ sejam iguais, isto é, $\lambda p_{111} = \mu p_4$. Se essas taxas de transição não fossem iguais e tivéssemos, por exemplo, $\lambda p_{111} > \mu p_4$, o sistema estaria em estado transiente com a cauda ainda em fase de

crescimento. Se o desequilíbrio fosse no sentido inverso, o hipercubo estaria em fase de crescimento em termos da massa de probabilidade. Em face dessa condição de equilíbrio, a equação em torno do estado {111} simplifica-se para

$$\{111\} \qquad \mu p_{111} = \lambda p_{011} + \lambda p_{101} + \lambda p_{110}, \qquad (8)$$

o que resulta em um sistema finito de equações (1)-(8), em termos das variáveis p_{000}, p_{001}, p_{010}, ..., p_{111}. Note que as demais probabilidades de equilíbrio de estados p_4, p_5, p_6, ..., podem ser facilmente obtidas substituindo-se os valores de p_{000}, p_{001}, p_{010}, ..., p_{111} nas equações de equilíbrio dos estados S_4, S_5, S_6, ...

2.3 Insuficiência das Equações de Equilíbrio

Se visualizarmos o sistema de equações (1)-(8) na forma $Ax = b$, constata-se que as equações de equilíbrio conduzem a um sistema com $b = 0$, isto é, os termos constantes das equações são todos nulos. Trata-se de um sistema homogêneo que tem uma solução trivial para $p_{ijk} = 0$, para $i,j,k = 0,1$. Além do mais, é um sistema indeterminado: pode-se arbitrar o valor de uma probabilidade particular e determinar as demais a partir dele. Existe uma explicação para tal indefinição: as equações impõem condição de equilíbrio em torno de cada estado do cubo, isto é, {000}, {001}, {010},... e {111}, mas nada especifica sobre a forma como a massa total de probabilidades se distribui entre esses e os estados da cauda do sistema S_4, S_5, S_6, …

A forma natural de levantar a indeterminação é a introdução de uma equação adicional de normalização: considerados todos os estados possíveis do sistema, a soma de suas probabilidades deve ser igual a um:

$$p_{000}+p_{001}+p_{010}+p_{100}+p_{011}+\ldots+p_{111}+p_4+p_5+\ldots = 1.$$

Como foi mencionado na seção 2.2, a condição de equilíbrio do sistema requer que as transições entre os estados {111} e {S_4} sejam iguais, isto é, $\lambda p_{111} = \mu p_4$. Em decorrência dessa condição, resulta que as transições entre os estados {S_k} e {S_{k+1}} para k=4,5,6… devem também ser iguais. Seja $\rho = \lambda/\mu < 1$; tem-se que:

$$p_4 = \rho p_{111};$$
$$p_5 = \rho p_4 = \rho^2 p_{111};$$
$$p_6 = \rho p_5 = \rho^3 p_{111};$$
$$\ldots$$

de modo que, a partir da progressão geométrica acima, chega-se a

$$p_{111}+p_4+p_5+\ldots = p_{111}/(1-\rho).$$

Assim, a equação de normalização das probabilidades pode ser escrita como

$$p_{000}+p_{001}+p_{010}+p_{100}+p_{011}+\ldots+ p_{111}/(1-\rho) = 1. \qquad (9)$$

Registre-se que, quando os servidores são homogêneos e têm a mesma taxa de atendimento ($\mu_1 = \mu_2 = \mu_3$), o modelo de hipercubo, no agregado, equivale ao modelo M/M/3. Isto permite que equações adicionais sejam incorporadas ao sistema, caso necessário. São elas:

$$p_{000} = \mathrm{P_{MM3}}\{S_0\}$$
$$p_{001}+p_{010}+p_{100} = \mathrm{P_{MM3}}\{S_1\}$$
$$p_{011}+p_{101}+p_{110} = \mathrm{P_{MM3}}\{S_2\}$$
$$p_{111} = \mathrm{P_{MM3}}\{S_3\}$$

onde $\mathrm{P_{MM3}}\{S_i\}$ representa a probabilidade de que um sistema M/M/3 se encontre no estado S_i. As equações acima, que resultam da equivalência, no agregado, entre o modelo hipercubo com servidores homogêneos e o modelo M/M/N, são denominadas de "hiperplanos" (LARSON, 1974).

2.4 Medidas de Desempenho

Com a distribuição de equilíbrio de estados (obtida com a solução do sistema (1)-(7) e (9)), podemos computar diversas medidas de desempenho interessantes para a análise do sistema. Por exemplo, a carga de trabalho

(*workload*) de cada servidor, isto é, a fração de tempo em que o servidor está ocupado, pode ser facilmente obtida somando-se as probabilidades dos estados em que este servidor está ocupado, ou seja:

$$\begin{aligned}\rho_1 &= p_{001} + p_{101} + p_{011} + p_{111} + p_Q \\ \rho_2 &= p_{010} + p_{110} + p_{011} + p_{111} + p_Q \\ \rho_3 &= p_{100} + p_{110} + p_{101} + p_{111} + p_Q,\end{aligned} \quad (10)$$

onde $p_Q = p_4 + p_5 + p_6 + ... = 1 - (p_{000} + p_{001} + p_{011} + ...+ p_{111})$ é a probabilidade de haver fila no sistema.

Outra medida de desempenho útil é a freqüência de despachos do servidor n para o átomo j:

$$f_{nj} = f_{nj}^{[1]} + f_{nj}^{[2]} = \frac{\lambda_j}{\lambda} \sum_{B \in E_{nj}} p_B + \frac{\lambda_j}{\lambda} p_s \frac{\mu_n}{\mu} \quad (11)$$

onde E_{nj} é o conjunto dos estados nos quais um chamado do átomo j é atendido pelo servidor n, e $p_s = p_Q + p_{111}$ é a probabilidade de saturação do sistema. O primeiro termo do lado direito de (11) ($f_{nj}^{[1]}$) corresponde à fração de todos os despachos do servidor n para o átomo j que não incorrem em qualquer tempo de espera em fila, enquanto o segundo termo ($f_{nj}^{[2]}$) corresponde à fração de todos os despachos do servidor n para o átomo j que incorrem em algum tempo de espera em fila.

Outras frações, como a fração de despachos do servidor n como *backup,* a fração de despachos de *backup* para o átomo j, e a fração total de despachos *backup* no sistema, podem ser fácilmente computadas (LARSON & ODONI, 1981).

Até agora não precisamos especificar os tempos médios de viagem entre os átomos i e j, τ_{ij}, para cada servidor n (veja hipótese 3). Esta matriz não precisa ser simétrica e pode refletir complicações na viagem devido a ruas e avenidas de mão simples, barreiras, condições de tráfego. Caso estes dados não possam ser obtidos empiricamente, pode-se utilizar conceitos de probabilidade geométrica para se estimar τ_{ij} para cada servidor.

Por simplicidade, nas expressões a seguir, admitimos que os tempos médios de serviço dos servidores são iguais a 1 (i.e., $\mu_1 = \mu_2 = \mu_3 = 1$). A partir de τ_{ij} , pode-se obter o tempo médio de viagem do sistema por:

$$T = \sum_{n=1}^{N} \sum_{j=1}^{N_A} f_{nj}^{[1]} t_{nj} + p_s T_Q$$

onde t_{nj} é o tempo médio de viagem que o servidor n, quando disponível, gasta para ir ao átomo j (note que t_{nj} é computado a partir de τ_{ij}), e $T_Q = \sum_{i=1}^{N_A} \sum_{j=1}^{N_A} \frac{\lambda_i \lambda_j}{\lambda^2} \tau_{ij}$ é o tempo médio de viagem até uma chamada que incorre em algum tempo de espera em fila (note que as razões λ_j/λ e λ_i/λ correspondem, respectivamente, à probabilidade de uma chamada que incorre em algum tempo de espera em fila ser gerada no átomo j, e à probabilidade de esta chamada ser atendida por um servidor localizado no átomo i). O tempo médio de viagem para cada átomo j é calculado por:

$$T_j = \frac{\sum_{n=1}^{N} f_{nj}^{[1]} t_{nj}}{\sum_{n=1}^{N} f_{nj}^{[1]}} (1 - p_s) + \sum_{i=1}^{N_A} \frac{\lambda_j}{\lambda} \tau_{ij} p_s$$

e o tempo médio de viagem de cada servidor n pode ser aproximado por:

$$TU_n = \frac{\sum_{j=1}^{N_A} f_{nj}^{[1]} t_{nj} + (T_Q p_s / N)}{\sum_{j=1}^{N_A} f_{nj}^{[1]} + (p_s / N)}. \quad (12)$$

Outras medidas de desempenho podem ser definidas. Para mais detalhes destas e de outras medidas de desempenho, o leitor pode consultar LARSON (1974) e LARSON & ODONI (1981).

2.5 Processo de Calibração dos Tempos Médios de Serviço

Em certos sistemas de atendimento emergencial, como certos casos de localização de bases

de ambulâncias, é provável que os tempos de viagem (tempo de viagem da base até o local do acidente, tempo de viagem do local do acidente até o hospital, tempo de viagem do hospital até a base) ocupem um papel relevante no tempo total de serviço. Neste caso, é importante que o tempo de serviço de cada unidade seja ajustado separadamente, para refletir os diferentes fatores geográficos afetando cada uma. Isto pode ser feito por meio de um processo de calibração dos tempos médios de serviço μ_n^{-1} de cada servidor n.

Considere novamente a expressão (12) para TU_n, o tempo médio de viagem para o servidor n. Se a soma de todos os tempos que compõem o tempo médio de serviço do servidor n (incluindo o valor computado TU_n) for diferente do valor μ_n^{-1} admitido inicialmente no modelo, então o modelo deve ser rodado novamente usando como parâmetro este "novo" tempo médio de serviço para o servidor n. E assim por diante, até que a diferença entre os tempos médios de serviço admitidos como parâmetros e os tempos médios de serviço computados pelo modelo sejam suficientemente próximos. A este procedimento iterativo dá-se o nome de processo de *calibração*. Experiência dos autores com diversos exemplos mostra que este procedimento costuma convergir em duas ou três iterações, para uma precisão razoável dos valores de μ_n^{-1}, embora uma prova que garanta esta convergência não tenha sido encontrada na literatura.

Na próxima seção discutimos métodos de solução exatos e aproximados para o modelo hipercubo.

3. A Solução do Modelo Hipercubo: Métodos Exato e Aproximado

Conforme apresentado na seção 2, a determinação das probabilidades de estado do modelo hipercubo requer a solução de um sistema linear construído a partir das equações de equilíbrio das taxas de transição. Para um sistema de N servidores, o sistema tem 2^N equações (note que, para o exemplo da seção 2 com N=3, o sistema tem 8 equações, isto é, (1)-(7) e (9)). Assim, mesmo para valores moderados de N, a solução do sistema deve ser cercada de alguns cuidados para assegurar-lhe viabilidade computacional. Depois de um certo limite, a solução do sistema de equações se torna inviável computacionalmente, sendo necessário recorrer a métodos aproximados.

Apresentamos nesta seção, em primeiro lugar, um método exato de solução do modelo hipercubo, no qual o sistema linear de equações é resolvido pelo método de Gauss-Siedel (para maiores detalhes sobre esse método veja por exemplo ATKINSON, 1978). Apresentamos em seguida o método aproximado de LARSON (1975).

3.1 O Método Exato

Iniciamos a apresentação do método exato de solução do modelo hipercubo a partir das equações de equilíbrio, tomando como base o exemplo simples de N = 3 servidores e N_A = 3 átomos da seção 2. Mostramos que as equações de equilíbrio do sistema se apresentam naturalmente em forma apropriada para o uso de métodos iterativos de solução e que, por envolverem matrizes de coeficientes esparsas, tais métodos apresentam a vantagem adicional de utilizarem exclusivamente os elementos não nulos da matriz dos coeficientes. Em conclusão, apresentamos alguns comentários sobre nossa limitada experiência computacional na solução de problemas teste.

A Solução do Sistema de Equações

Consideremos novamente o sistema de equações de equilíbrio (1)-(7) e (9), desenvolvido na seção 2. Para escolher um método apropriado para resolver este sistema, observe-se que a equação de equilíbrio (1) em torno do estado {000} permite escrever de imediato:

$$p_{000} = (\mu_1 p_{001} + \mu_2 p_{010} + \mu_3 p_{100})/\lambda$$

Esta forma, em que p_{000} fica expressa em termos das demais probabilidades, sugere o uso de um método iterativo de solução, definindo-se a iteração

$$p_{000}(k) = [\mu_1 p_{001}(k\text{-}1) + \mu_2 p_{010}(k\text{-}1) + \\ + \mu_3 p_{100}(k\text{-}1)]/\lambda$$

onde $p_{ijk}(k)$ representa o valor obtido para p_{ijk} na k-ésima iteração. De forma análoga, a equação de equilíbrio (2) em torno do estado {001} conduz à iteração

$$p_{001}(k) = [\lambda_1 p_{000}(k\text{-}1) + \mu_2 p_{011}(k\text{-}1) + \\ + \mu_3 p_{101}(k\text{-}1)]/(\lambda+\mu_1)$$

para p_{001}. Esta iteração caracteriza o método de Gauss-Jacobi (veja ATKINSON, 1978), em que, a cada iteração, se utiliza os valores das variáveis obtidos na iteração anterior. Pode-se no entanto utilizar os valores mais recentes das variáveis, no sentido que, por exemplo, no momento de determinar $p_{001}(k)$, o valor mais atualizado de p_{000} é $p_{000}(k)$, e não $p_{000}(k\text{-}1)$. Pode-se utilizar:

$$p_{001}(k) = [\lambda_1 p_{000}(k) + \mu_2 p_{011}(k\text{-}1) + \\ + \mu_3 p_{101}(k\text{-}1)]/(\lambda+\mu_1), \qquad (13)$$

obtendo-se então a iteração Gauss-Siedel.

Observe-se que o somatório do segundo membro da expressão (13) tem três termos. Isto reflete o fato que, no espaço tridimensional, a probabilidade associada ao estado representado pelo vértice {001} do cubo depende apenas das probabilidades associadas aos estados representados pelos vértices que são adjacentes a {001}. Este fato é verdadeiro também para N genérico: como os estados do sistema que se comunicam estão sempre em vértices adjacentes do hipercubo, qualquer equação de equilíbrio terá sempre N+1 termos. Assim, embora a matriz dos coeficientes seja de dimensão 2^Nx2^N, cada linha da matriz tem, além do elemento diagonal, apenas N elementos não nulos.

Dessa forma um método iterativo, no caso o método de Gauss-Siedel baseado na expressão (13), além de ser uma escolha natural pela forma das equações de equilíbrio, apresenta a vantagem adicional de ser apropriado pela natureza da matriz de coeficientes do sistema, que é bastante esparsa para valores grandes de N.

Os métodos iterativos de solução de sistemas lineares de equações são métodos aproximados por natureza, no sentido de que sua implementação requer um critério de parada baseado no erro absoluto máximo tolerável. Além do mais, os métodos iterativos podem não convergir. No caso do método de Gauss-Siedel (assim como no de Gauss-Jacobi), sabe-se que a condição suficiente de convergência é de que a matriz dos coeficientes seja diagonalmente dominante, isto é, que o elemento diagonal de uma linha seja, em módulo, maior que a soma dos módulos dos demais elementos da mesma linha. A matriz dos coeficientes do sistema de equações do modelo hipercubo não satisfaz esta condição, de modo que não existe garantia formal de convergência do método de Gauss-Siedel. No entanto, existem duas considerações que parecem justificar a conjectura de que o método seja convergente. Em primeiro lugar, a natureza do problema: trata-se de um sistema de equações em que as incógnitas são probabilidades que são normalizadas para totalizar um a cada iteração. Em segundo lugar, a nossa experiência computacional envolvendo algumas dezenas de problemas teste não revelou uma única instância em que o método não convergisse.

Experiência Computacional

O método de Gauss-Siedel foi testado para alguns problemas gerados aleatoriamente em sistemas com até 17 servidores, tendo como dados de entrada o número de átomos N_A, o número de servidores N e a utilização média $\rho < 1$, o gerador:

- utilizou amostras da distribuição uniforme entre 0 e 1 como λ_j, j=1,..,N_A;
- utilizou amostras da distribuição uniforme entre 0,2 e 1,2 como μ_i, i=1..,N, e, numa

segunda passada, normalizou os μ_i para produzir ρ desejado e

- cada linha da matriz de despachos foi gerada como uma amostra aleatória sem reposição de N elementos de [1, 2, 3, …, N].

Usando 2^{-N} como valores iniciais das probabilidades, constatou-se grandes oscilações dos erros nas primeiras iterações. Entretanto, após 15 a 20 iterações, o erros máximos mostraram redução a taxas praticamente constantes nos problemas testados, comportamento típico de sistemas convergentes.

Deve-se atribuir as grandes oscilações dos erros no início do processo à inadequação da escolha de 2^{-N} como valor inicial das probabilidades. Por outro lado, o comportamento do processo após a fase inicial de instabilidade parece indicar que o sistema em si é bem comportado, não devendo apresentar problemas de convergência do método de Gauss-Siedel em sua solução. Coloca-se então o problema de definir valores iniciais mais apropriados para aplicar o método iterativo de solução. Os hiperplanos de Larson oferecem uma boa alternativa nesse sentido.

Os hiperplanos resultam do fato de que, para sistemas com servidores homogêneos (i.e., $\mu_1 = \mu_2 = \ldots = \mu_N$), as probabilidades dos estados do hipercubo são, no agregado, iguais às probabilidades dos estados do sistema M/M/N. Assim, para a estado vazio de um sistema de três servidores teríamos:

$$p_{000} = \mathrm{P}_{\mathrm{MM3}}\{S_0\}$$

Para sistemas com servidores distinguíveis esta condição não é verdadeira, mas parece razoável supor-se que $\mathrm{P}_{\mathrm{MM3}}\{S_0\}$ seja um valor adequado para inicializar p_{000}:

$$p_{000}(0) = \mathrm{P}_{\mathrm{MM3}}\{S_0\}$$

O hiperplano relativo ao estado em que um usuário se encontra presente no sistema é dado pela equação:

$$p_{001} + p_{010} + p_{100} = \mathrm{P}_{\mathrm{MM3}}\{S_1\}$$

Esta equação, não obstante aplicar-se somente a sistemas com servidores homogêneos, permite obter valores iniciais mais apropriados das probabilidades na resolução de equações associadas a sistemas com servidores distinguíveis pelo método de Gauss-Siedel, distribuindo-se a massa do hiperplano igualmente entre os estados que definem o hiperplano:

$$p_{001}(0) = p_{010}(0) = p_{100}(0) = \mathrm{P}_{\mathrm{MM3}}\{S_1\}/3$$

De forma semelhante, os valores iniciais das demais probabilidades podem ser determinadas a partir dos hiperplanos associados aos estados com dois e três usuários presentes no sistema:

$$p_{011}(0) = p_{101}(0) = p_{110}(0) = \mathrm{P}_{\mathrm{MM3}}\{S_2\}/3$$

$$p_{111}(0) = \mathrm{P}_{\mathrm{MM3}}\{S_3\}$$

Os valores iniciais das probabilidades assim definidos tornou o método de Gauss-Siedel estável no sentido de que os erros máximos mostram-se decrescentes a partir das iterações iniciais em todos os problemas testados.

Usando como critério de convergência um erro relativo máximo menor que 0,001%, o pior caso foi para $N = 17$ e $\rho = 0.9$, com tempo de CPU de 58 segundos para resolver o sistema de equações (os testes foram rodados em micro Pentium II com *clock* de 333Mhz e 64Mb de RAM). Em geral, para ρ menores, a convergência é mais rápida: para $\rho = 0.5$ convergências foram obtidas entre 25 e 30 segundos de processamento.

Na realidade, para problemas maiores, o elemento crítico passou a ser o tempo necessário para gerar os coeficientes do sistema de equações. Para problemas com 15, 16 e 17 servidores, tais tempos foram de 32, 73 e 158 segundos, respectivamente. Embora pouco cuidado tenha sido tomado na elaboração do código para gerar os coeficientes, a experiência deixou a impressão de que o uso do método

exato parece pouco operacional na implementação de métodos de substituição de vértices para problemas de localização, em que a avaliação da troca de um par de vértices requer a solução de um modelo hipercubo (veja seção 4.4 abaixo).

3.2 O Método Aproximado de Larson

Resolver o modelo hipercubo significa obter os elementos necessários para determinar suas características operacionais básicas. Assim, pelo método exato obtém-se as probabilidades associadas aos estados descritos pelos vértices do hipercubo. Tais probabilidades são, por sua vez, utilizadas para determinar as medidas de desempenho do sistema (veja seção 2.4).

Para um sistema com N servidores, o método exato requer a solução de um sistema linear de 2^N equações. Assim, mesmo valores moderados de N podem envolver sistemas com portes, senão intratáveis, pouco operacionais para o uso do método exato. Como alternativa, LARSON (1975) desenvolveu um método aproximado, que envolve um sistema não linear de equações que tem como incógnitas as N taxas de ocupação dos servidores, e não as 2^N probabilidades de estado do método exato. O desenvolvimento do método se baseia na hipótese de que todos os servidores têm uma taxa de serviço comum μ.

Para descrever o método aproximado de Larson, apresentamos inicialmente a maneira de calcular as freqüências de despachos a partir das taxas de ocupação dos servidores. A seguir derivamos uma expressão para a taxa total de despachos de um servidor particular. Mostramos em seguida a relação que existe entre a taxa total de despachos de um servidor e sua taxa média de ocupação (carga de trabalho), relação esta que permite derivar a fórmula cujo uso iterativo permite determinar numericamente as taxas médias de ocupação dos servidores. Finalmente, após descrever a derivação do fator de correção necessário para levar em conta a não independência dos servidores, concluímos comentando em que reside precisamente o caráter aproximado do método.

Freqüências de Despachos

A freqüência de despachos do servidor n para o átomo j, f_{nj}, conforme definida em (11), é a fração de todas as chamadas que resulta no envio da unidade n para o átomo geográfico j, com $\Sigma_{n,j} f_{n,j}=1$. Esta freqüência é constituída de duas parcelas: a primeira, $f_{nj}^{[1]}$, relativa a chamadas que são atendidas sem espera e a segunda, $f_{nj}^{[2]}$, relativa a chamadas sujeitas a espera. Consideremos o exemplo de três servidores da seção 2, em que o átomo 1 tem o servidor 1 como servidor preferencial e os servidores 2 e 3 como primeiro e segundo *backup*, respectivamente. A freqüência de despachos $f_{11}^{[1]}$ é dada pelo produto das probabilidades de que: (i) uma dada chamada se origina no átomo 1 e (ii) o servidor 1 está livre. Se as probabilidades de estado do sistema tivessem sido determinadas pelo método exato, teríamos (veja expressão (11)):

$$f_{11}^{[1]} = (\lambda_1/\lambda)(p_{000} + p_{100} + p_{010} + p_{110}).$$

Alternativamente, supondo-se desconhecidas as probabilidades de estado, poderíamos rescrever a expressão acima como (veja expressão (10)):

$$f_{11}^{[1]} = (\lambda_1/\lambda)(1-\rho_1)$$

onde ρ_1, conforme seção 2.4, representa a taxa média de ocupação do servidor 1 (i.e., sua carga de trabalho).

A parcela $f_{nj}^{[2]}$ em (11) relativa a chamadas com espera independe das probabilidades de estado ou taxas individuais de ocupação dos servidores, sendo aqui apresentada apenas para tornar completa a descrição da forma de determinar as freqüências de despachos. A fração de todas as chamadas que resulta no envio da unidade 1 para o átomo 1 com espera é obtida pelo produto das probabilidades de que: (i) uma dada chamada se origina no átomo 1, (ii) o sistema está saturado e (iii) o servidor 1 é escolhido para atender a chamada:

$$f_{11}^{[2]} = (\lambda_1/\lambda)P_s(1/3) \qquad (14)$$

onde P_s representa a probabilidade de saturação do sistema, conforme seção 2.4. Note em (14) que, no cálculo da probabilidade (iii), supõe-se que a disciplina de atendimento das chamadas em espera independe da localização do átomo de origem das chamadas e do servidor escolhido para atendê-las. Assim, como se trata de um sistema com servidores homogêneos, todos os servidores têm a mesma probabilidade de serem escolhidos para atender uma chamada em espera (veja parcela 1/3 do lado direito de (14)). Observe-se que, em virtude da equiprobabilidade da escolha, esta parcela da freqüência de despachos, para um dado átomo geográfico, independe do servidor escolhido.

Considere-se em seguida a freqüência $f_{21}^{[1]}$. De vez que o servidor 2 é o primeiro *backup* do servidor 1, servidor preferencial do átomo 1, esta freqüência consiste do produto das probabilidades de que: (i) uma dada chamada se origina no átomo 1, e (ii) o servidor 1 está ocupado e o servidor 2 está livre:

$$f_{21}^{[1]} = (\lambda_1/\lambda)\text{Prob}\{B_1F_2\} \qquad (15)$$

onde B_1 representa o evento de o servidor 1 estar ocupado e F_2 o evento de o servidor 2 estar livre. Se as probabilidades de estado do sistema tivessem sido determinados pelo método exato, teríamos:

$$f_{21}^{[1]} = (\lambda_1/\lambda)(p_{001}+p_{101})$$

Para derivar uma forma alternativa para calcular $f_{21}^{[1]}$ em (15), observe-se que, se os servidores operassem independentemente uns dos outros, a probabilidade do evento $\{B_1F_2\}$ seria igual a $\rho_1(1-\rho_2)$.

No seu método aproximado, LARSON propõe a utilização de um fator de correção $Q(N,\rho,k)$ para levar em conta a não independência entre k+1 servidores, obtendo-se

$$\text{Prob}\{B_1F_2\} = Q(N, \rho,1)\ \rho_1(1-\rho_2) \qquad (16)$$

onde $\rho=\lambda/\mu$ representa a taxa de ocupação média do sistema. Usando (16), a freqüência de despachos do servidor 2 ao átomo 1 sem espera em (15) pode ser escrita como

$$f_{21}^{[1]} = (\lambda_1/\lambda)Q(N, \rho,1)\ \rho_1(1-\rho_2)$$

Considerações análogas conduzem à expressão relativa à freqüência de despachos do servidor 3 ao átomo 1 sem espera:

$$f_{31}^{[1]} = (\lambda_1/\lambda)Q(N, \rho,2)\ \rho_1\rho_2\ (1-\rho_3)$$

Verifica-se assim que as freqüências de despachos de servidores a átomos em um modelo hipercubo podem ser determinadas sem o uso das probabilidades dos estados do sistema, desde que se conheça as taxas médias de ocupação dos servidores e o fator de correção para levar em conta a não independência dos servidores.

Taxa Total de Despachos de um Servidor

Conforme visto acima, para resolver o modelo hipercubo não utilizando as probabilidades de estado do sistema, é necessário construir uma equação que permita determinar numericamente as taxas médias de ocupação dos servidores. Tal equação pode ser construída a partir da expressão que descreve a taxa total de despachos de um dado servidor. Seja então R_n a taxa total de despachos do servidor n, isto é, o número de vezes que o servidor n é despachado para atendimento na unidade de tempo.

A taxa R_n é constituída de N+1 parcelas. As primeiras N parcelas referem-se à situação em que o servidor n encontra-se livre para iniciar imediatamente o atendimento de uma chamada. Estas parcelas compreendem os despachos do servidor n na qualidade de *j-ésimo* servidor preferencial para j=1,2,…,N. Um elemento genérico deste grupo é constituído pelas taxas de chegadas dos átomos para os quais o servidor n é o *j-ésimo* servidor preferencial, multiplicada pela fração de tempo em que os primeiros (j-1) servidores preferenciais destes átomos estão ocupados e o servidor n está livre. A última

parcela é constituída pela fração das chamadas sujeitas à espera que são rateadas igualmente entre todos os servidores, sendo igual a $\lambda P_s/N$.

Definindo G_n^j como o conjunto de átomos para os quais o servidor n é o *j-ésimo* servidor preferencial e m_{aj} como o *j-ésimo* servidor preferencial do átomo a, tem-se:

$$R_n = \sum_{a\in G_n^1}\lambda_a(1-\rho_n) + \sum_{a\in G_n^2}\lambda_a \text{Prob}\{B_{m_{a1}}F_n\} + \\ + \sum_{a\in G_n^3}\lambda_a \text{Prob}\{B_{m_{a1}}B_{m_{a2}}F_n\} + \ldots + \\ + \sum_{a\in G_n^N}\lambda_a \text{Prob}\{B_{m_{a1}}B_{m_{a2}}\ldots B_{m_{a(N-1)}}F_n\} + \lambda_D,$$

onde $\lambda_D = \lambda P_s/N$.

Valendo-se do recurso de representar as probabilidades conjuntas que figuram na expressão acima como as probabilidades aplicáveis sob a hipótese de independência dos servidores corrigidas pelo fator Q, tem-se (veja expressão (16)):

$$R_n = \sum_{a\in G_n^1}\lambda_a(1-\rho_n) + \sum_{a\in G_n^2}\lambda_a Q(N,\rho,1)\rho_{m_{a1}}(1-\rho_n) + \\ + \sum_{a\in G_n^3}\lambda_a Q(N,\rho,2)\rho_{m_{a1}}\rho_{m_{a2}}(1-\rho_n) + \ldots + \\ + \sum_{a\in G_n^N}\lambda_a Q(N,\rho,N-1)\rho_{m_{a1}}\rho_{m_{a2}}\ldots\rho_{m_{a(N-1)}}(1-\rho_n) + \lambda_D. \quad (17)$$

Note em (17) que R_n, a taxa total de despachos do servidor n, fica expressa em termos das taxas de ocupação dos servidores. Como está, esta equação ainda não é suficiente para a determinação das taxas de ocupação dos servidores. Na verdade, necessita-se obter a relação que existe entre R_n e ρ_n para obter uma equação para ρ_n.

Relação entre a Taxa Total de Despachos e a Taxa de Ocupação de um Servidor

Para obter a relação entre a taxa total de despachos e a taxa de ocupação do servidor n, considere-se um intervalo de tempo T. Uma vez que o servidor n é despachado R_n vezes em cada unidade de tempo, o número total de despachos do servidor n no intervalo de tempo T será R_nT. Sendo μ a taxa de atendimento do servidor, R_nT chamadas demandarão um tempo total de atendimento de $\mu^{-1}R_nT$ unidades de tempo. Assim, $\mu^{-1}R_nT/T$ ou $\mu^{-1}R_n$ representa a ocupação média do servidor n em T. Desde que T seja suficientemente grande, $\mu^{-1}R_n$ pode ser utilizado para representar a taxa de ocupação do servidor n no estado estacionário, isto é, $\rho_n = \mu^{-1}R_n$. Além do mais, se adotarmos como tempo médio de atendimento dos servidores a unidade de tempo (i.e., $\mu^{-1}=1$), tem-se que a taxa total de despachos de um servidor é igual à sua taxa média de ocupação.

Então, valendo-se da relação $R_n = \rho_n$, a equação (17) pode ser utilizada para escrever:

$$\frac{\rho_n - \lambda_D}{1-\rho_n} = \sum_{a\in G_n^1}\lambda_a + \sum_{a\in G_n^2}\lambda_a Q(N,\rho,1)\rho_{m_{a1}} + \\ + \sum_{a\in G_n^3}\lambda_a Q(N,\rho,2)\rho_{m_{a1}}\rho_{m_{a2}} + \ldots + \\ + \sum_{a\in G_n^N}\lambda_a Q(N,\rho,N-1)\rho_{m_{a1}}\rho_{m_{a2}}\ldots\rho_{m_{a(N-1)}}.$$

Definindo

$$R_n^F = \sum_{a\in G_n^1}\lambda_a + \sum_{a\in G_n^2}\lambda_a Q(N,\rho,1)\rho_{m_{a1}} + \\ + \sum_{a\in G_n^3}\lambda_a Q(N,\rho,2)\rho_{m_{a1}}\rho_{m_{a2}} + \ldots + \\ + \sum_{a\in G_n^N}\lambda_a Q(N,\rho,N-1)\rho_{m_{a1}}\rho_{m_{a2}}\ldots\rho_{m_{a(N-1)}},$$

tem-se:

$$\rho_n = \frac{R_n^F + \lambda_D}{1+R_n^F}.$$

A equação acima expressa a taxa de ocupação do servidor n em termos das taxas de ocupação de todos os demais servidores. Trata-se de uma forma particularmente apropriada para uso em métodos iterativos de resolução numérica em que, a cada nova iteração, um valor melhorado de uma variável é obtido a partir dos valores mais recentes de todas as demais variáveis.

A implementação de tal método fica então dependente de uma expressão para determinar o fator de correção Q.

O Fator de Correção *Q(N,ρ, j)*

Como vimos acima, a necessidade do fator de correção Q ocorre quando se busca expressar a probabilidade associada a um evento da forma $\{B_1B_2...B_jF_{j+1}\}$ de que uma chamada encontre seus primeiros j servidores preferenciais ocupados e o *(j+1)-ésimo* servidor preferencial livre. Argumentou-se então que, se os servidores operassem independentemente uns dos outros, a probabilidade desejada seria dada por $\rho_1\rho_2...\rho_j(1-\rho_{j+1})$. Mesmo que na realidade os servidores não sejam independentes, esta expressão poderia ser utilizada, desde que se introduza um fator de correção Q capaz de levar em conta a não independência entre servidores, tendo-se então:

$$\text{Prob}\{B_1B_2...B_jF_{j+1}\} = Q(N,\rho,j)\ \rho_1\rho_2...\rho_j(1-\rho_{j+1}) \quad (18)$$

Como a dependência entre servidores é de natureza complexa e obviamente condicionada pela política de despachos do modelo hipercubo, LARSON (1975) conjeturou que um fator de correção similar derivado a partir de pressupostos mais genéricos poderia aproximar satisfatoriamente os resultados necessários para descrever as características de um modelo hipercubo.

A estrutura mais genérica que LARSON utilizou tem como base um modelo de fila M/M/N. Em relação a este modelo, imagina-se um experimento que consiste na amostragem aleatória sem reposição de servidores, e define-se o evento $\{B_1B_2...B_jF_{j+1}\}$ de que os primeiros j servidores escolhidos estejam ocupados e o *(j+1)-ésimo* servidor amostrado esteja livre.

Utilizando-se as leis da probabilidade condicional, temos:

$$\text{Prob}\{B_1B_2...B_jF_{j+1}\} = \Sigma_k\ \text{Prob}\{B_1B_2...B_jF_{j+1}|S_k\}.P_k \quad (19)$$

onde S_k representa o estado em que k usuários se encontram presentes no sistema, e P_k a probabilidade associada a S_k. Para que a probabilidade do evento $\{B_1B_2...B_jF_{j+1}\}$ seja não nula, é necessário que existam pelo menos j servidores ocupados e pelo menos um servidor livre. Por esta razão, o somatório em (19) deve estender-se para k variando de j a N-1.

A probabilidade condicional que figura na expressão (19) pode ser escrita como

$$\text{Prob}\{B_1B_2...B_jF_{j+1}|S_k\} = \text{Prob}\{B_1B_2...B_jF_{j+1}S_k\}/P_k \quad (20)$$

e a probabilidade conjunta que figura na expressão (20) pode, por sua vez, ser escrita como:

$$\begin{aligned}&\text{Prob}\{B_1B_2...B_jF_{j+1}S_k\} = \\ &= \text{Prob}\{F_{j+1}|B_1B_2...B_jS_k\}.\text{Prob}\{B_j|B_1B_2...B_{j-1}S_k\}... \\ &\quad ...\text{Prob}\{B_2|B_1S_k\}\ \text{Prob}\{B_1|S_K\}.P_k\end{aligned}$$

de modo que, substituindo esta equação em (20), temos que:

$$\begin{aligned}&\text{Prob}\{B_1B_2...B_jF_{j+1}|S_k\} = \\ &= \text{Prob}\{F_{j+1}|B_1B_2...B_jS_k\}.\text{Prob}\{B_j|B_1B_2...B_{j-1}S_k\}... \\ &\quad ...\text{Prob}\{B_2|B_1S_k\}.\text{Prob}\{B_1|S_k\} \quad (21)\end{aligned}$$

As probabilidades condicionais do segundo membro da expressão (21) podem ser facilmente avaliadas. Começando pela última, observe-se que $\{B_1|S_k\}$ representa o evento de selecionar um servidor ocupado dado que k dentre N servidores estão ocupados. A probabilidade associada é dada por

$$\text{Prob}\{B_1|S_k\} = k/N$$

O termo seguinte em (21) envolve o evento $\{B_2|B_1S_k\}$. Representa o evento de que o segundo servidor selecionado esteja ocupado, dado que o primeiro servidor selecionado está ocupado e k usuários estão presentes no sistema. Como se trata de amostragem sem reposição e o primeiro servidor está ocupado, no momento da seleção

do segundo servidor existem ainda N-1 servidores passíveis de seleção, dos quais k-1 estão ocupados. Portanto

$$\text{Prob}\{B_2|B_1S_k\} = (k\text{-}1)/(N\text{-}1)$$

Argumentações análogas permitem determinar as demais probabilidades. Em particular, para os dois primeiros termos em (21) tem-se

$$\text{Prob}\{B_j|B_1B_2...B_{j-1}S_k\} = [k\text{-}(j\text{-}1)]/[N\text{-}(j\text{-}1)]$$

$$\text{Prob}\{F_{j+1}|B_1B_2...B_jS_k\} = (N\text{-}k)/(N\text{-}j)$$

Substituindo-se as expressões acima em (21), obtém-se:

$$\text{Prob}\{B_1B_2...B_jF_{j+1} \mid S_k\} =$$

$$= \frac{k}{N}\frac{k-1}{N-1}...\frac{k-j+1}{N-j+1}\frac{\text{N - k}}{\text{N - j}} =$$

$$= (N-k)\frac{k(k-1)...(k-j+1)}{N(N-1)...(N-j)} =$$

$$= (N-k)\frac{\frac{k!}{(k-j)!}}{\frac{N!}{(N-j-1)!}} =$$

$$= \frac{(N-k)\,k!\,(N-j-1)!}{(k-j)!\,N!}.$$

Ponderando-se a probabilidade condicional acima sobre a distribuição de S_k, obtém-se a probabilidade incondicional desejada. Como S_k representa os estados de um sistema M/M/N, sua distribuição é dada por

$$P_k = \text{Prob}\{S_k\} = \frac{N^k\rho^kP_0}{k!},$$

para k = 1,..,N-1, sendo:

$$P_0 = \frac{1}{\sum_{i=0}^{N-1}\frac{N^i\rho^i}{i!} + \frac{N^N\rho^N}{N!(1-\rho)}}.$$

Após alguma manipulação algébrica com as expressões (18), (19) e as acima, obtém-se:

$$Q(N,\rho,j) = \frac{\sum_{k=j}^{N-1}\{\frac{(N-j-1)!(N-k)}{(k-j)!}\}\frac{N^k}{N!}\rho^{k-j}}{(1-\rho)[\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i] + \frac{N^N\rho^N}{N!}}.$$

Com a derivação do fator de correção Q, completa-se a descrição do método aproximado de solução do modelo hipercubo. Em conclusão, aponta-se a seguir as características que tornam o método aproximado.

O Caráter Aproximado do Método

São dois os aspectos do desenvolvimento do método de Larson que lhe conferem o caráter de ser aproximado. O primeiro é a hipótese de que os servidores são homogêneos: todos têm a mesma taxa de serviço. Nas palavras do próprio autor, "…a aproximação decorre do fato de que no contexto de um sistema urbano de serviços, os servidores são distinguíveis e têm diferentes características de desempenho." O segundo aspecto é o fato de que a derivação do fator Q baseia-se em um processo de amostragem aleatória de servidores em um sistema M/M/N, enquanto no modelo hipercubo a seleção dos servidores se dá na seqüência estabelecida pela política de despachos.

A conjectura é de que, apesar disso, o mesmo fator pode ser usado como uma aproximação em um modelo hipercubo com servidores distinguíveis, particularmente em sistemas em que: (i) as taxas de ocupação dos servidores não sejam muito diferentes e (ii) muitos vetores de preferência de despachos simulem, no conjunto, uma seleção aleatória de servidores.

Os resultados apresentados por LARSON (1975), com desvios da ordem de 1 a 2% em relação aos resultados do método exato, comprovam a utilidade do método aproximado.

4. Problemas de Localização Probabilísticos

Problemas de localização probabilísticos se ocupam, de uma maneira geral, da natureza estocástica de sistemas do mundo real. Nesses

sistemas um dado número de parâmetros, tais como por exemplo tempos de viagem, custos de construção, localização de clientes, quantidades demandadas e disponibilidade de servidores, são variáveis aleatórias. O objetivo é localizar facilidades/servidores que otimizem a medida de utilidade escolhida de forma robusta, para uma gama de valores dos parâmetros considerados. Os problemas estocásticos consideram explicitamente as distribuições probabilísticas das variáveis aleatórias modeladas. Segundo OWEN & DASKIN (1998), alguns autores incorporam essas distribuições em formulações padrão de programação matemática, enquanto outros as utilizam em um enfoque de teoria das filas. Nosso interesse neste trabalho se concentra no enfoque de teoria das filas.

Uma revisão bibliográfica detalhada de problemas de localização probabilísticos é apresentada por OWEN & DASKIN (1998). SWERSEY (1994) examina também alguns modelos probabilísticos. Destes nos interessam particularmente os modelos que tratam como variável aleatória a disponibilidade dos servidores, que tem bastante importância na localização de serviços de emergência. Na modelagem probabilística de serviços de emergência, algumas hipóteses simplificadoras permitem o uso da programação matemática, através, por exemplo, da definição de *restrições de oportunidade* (*chance-constrained programming*). Situações em que hipóteses simplificadoras não são aplicáveis conduzem, no entanto, ao tratamento desses problemas através da modelagem de filas espacialmente distribuídas, na qual assume fundamental importância o modelo hipercubo de Larson discutido nas seções 2 e 3 acima.

4.1 A Localização de Serviços de Emergência

Na localização de serviços de emergência busca-se em geral prover *cobertura* a áreas de demanda. A noção de *cobertura* implica na definição de uma *distância (tempo) de serviço*, que é a *distância (tempo) crítica* além da qual a área de demanda é considerada *não coberta*. Uma área de demanda é portanto considerada *coberta* se está a menos da distância crítica de pelo menos uma das facilidades (servidores) existentes, independentemente de a facilidade (ou servidor) estar ou não disponível quando o serviço é solicitado.

Considere por exemplo o problema de localizar serviços de atendimento de emergência por ambulâncias ou por estações do corpo de bombeiros em uma dada região, de tal modo que toda a população da região esteja a menos de, digamos, 10 quilômetros de pelo menos uma das facilidades. Neste caso 10 quilômetros definem a *distância crítica* e o problema consiste na determinação do número mínimo de facilidades e de sua localização na região em consideração, de tal forma que cada área de demanda esteja a menos de 10 quilômetros de pelo menos uma das facilidades localizadas.

O mais simples dos modelos matemáticos existentes para problemas de localização com restrições de cobertura, que resolve o problema acima, é o do *Problema de Localização para a Cobertura de Conjuntos (PLCC)*. Um problema a ele relacionado é o *problema de localização dos p-centros*, que busca a localização de p facilidades de tal forma que a distância máxima de qualquer área de demanda à facilidade mais próxima seja a mínima possível.

Muitas vezes atingir a cobertura total para uma dada região (conforme acima exemplificado), para uma distância crítica predefinida, pode tornar-se inviável do ponto de vista econômico, no sentido que o número de facilidades necessárias pode não ser compatível com os recursos disponíveis. Usualmente busca-se uma solução de compromisso, que proporcione níveis de cobertura aceitáveis e seja financeiramente mais acessível. *O Problema de Localização de Máxima Cobertura (PLMC)* foi desenvolvido com este propósito. Neste caso o objetivo é localizar um número pré-especificado de facilidades (digamos p facilidades), compatível com os recursos disponíveis, tal que a máxima população possível de uma dada região seja coberta a menos de uma distância crítica S predefinida.

O PLMC foi inicialmente estudado por CHURCH & REVELLE (1974). Matematicamente ele pode ser definido da seguinte forma. Seja S a *distância de serviço* definida para o problema e sejam $J = \{1,2,...,m\}$ o conjunto de áreas de demanda, $I = \{1,2,...,n\}$ o conjunto de locais em potencial onde facilidades podem ser localizadas, f_j a população da área de demanda j, $a_{ij} = 1$ se a área de demanda j puder ser coberta por uma facilidade localizada em $i \in I$ a menos da distância de serviço S ($a_{ij} = 0$ caso contrário), e p o número de facilidades a serem localizadas. Convencionemos por outro lado $x_j = 1$ se a área de demanda j for coberta ($x_j = 0$ caso contrário); $y_i = 1$ significa que uma facilidade deve ser localizada em $i \in I$ ($y_i = 0$ caso contrário). A formulação matemática de PLMC é dada por:

$$\text{Maximize} \quad Z = \sum_{j \in J} f_j x_j \tag{22}$$

$$\text{sujeito a} \quad \sum_{i \in I} a_{ij} y_i - x_j \geq 0, \; j \in J, \tag{23}$$

$$\sum_{i \in I} y_i = p, \tag{24}$$

$$x_j \in \{0,1\}, \; j \in J, \tag{25}$$

$$y_i \in \{0,1\}, \; i \in I. \tag{26}$$

Na formulação acima a função objetivo busca maximizar a população total coberta. As restrições (23) asseguram que uma área de demanda $j \in J$ está coberta se existe pelo menos uma facilidade a menos da distância S da mesma. A restrição (24) limita o número de facilidades na solução a p. Finalmente, as restrições (25)-(26) definem a natureza binária das variáveis de decisão.

Uma desvantagem dos modelos determinísticos de localização de serviços de emergência acima definidos é que eles partem da hipótese que as facilidades (servidores) estão disponíveis quando solicitadas, o que nem sempre é verdadeiro em aplicações práticas. Em sistemas não congestionados, com pouca demanda, a hipótese é razoável, mas em sistemas congestionados, nos quais chamadas freqüentes mantêm por exemplo ambulâncias na rua 20% a 30% do tempo, a hipótese é totalmente injustificada. O congestionamento em serviços de atendimento de emergência, que pode causar a indisponibilidade de um servidor a menos da distância crítica quando solicitado, motivou o desenvolvimento dos modelos de localização com cobertura adicional e, posteriormente, dos modelos de cobertura probabilísticos.

4.2 Modelos de Localização com Cobertura Adicional

Partindo da possibilidade que em sistemas congestionados o primeiro (e em muitos casos o único) servidor disponível para atender uma dada área de cobertura esteja ocupado quando solicitado, foram desenvolvidos diversos modelos com cobertura adicional. Exemplos desses são os modelos BACOP1 e BACOP2 (representando respectivamente os *Backup Coverage Problems 1 & 2*) desenvolvidos por HOGAN & REVELLE (1986). Em BACOP1 os autores enfatizam a importância de pelo menos um servidor adicional para cada área de demanda. Este modelo, tal como em PLCC, exige que cada área de demanda seja sempre coberta; além disso sua função objetivo busca maximizar a população coberta por pelo menos um servidor adicional.

Seja a variável $u_j \in \{0,1\}$ tal que seu valor é igual a 1 se existirem pelo menos dois servidores disponíveis a menos de S da área de demanda j ($u_j = 0$ caso contrário); BACOP1 pode ser formulado matematicamente da seguinte forma:

$$\text{Maximize} \quad Z = \sum_{j \in J} f_j u_j, \tag{27}$$

$$\text{sujeito a} \quad u_j \leq \sum_{i \in I} a_{ij} y_i - 1, \; j \in J, \tag{28}$$

$$\sum_{i \in I} y_i = p, \tag{29}$$

$$u_j \in \{0,1\}, \; j \in J, \tag{30}$$

$$y_i \in \{0,1\}, \; i \in I. \tag{31}$$

Nessa formulação a função objetivo busca maximizar a população com uma segunda cobertura a menos de *S*. As restrições (28) asseguram que uma área de demanda $j \in J$ possui dupla cobertura se existirem pelo menos dois servidores a menos de *S* da mesma. É importante notar que como este modelo exige primeira cobertura para todas as áreas de demanda e existe uma restrição no número máximo de facilidades na solução (restrição (29)), o valor de *p* deve ser previamente calculado através do modelo PLCC. Este valor de *p* previamente calculado assegura pelo menos uma cobertura para todas as áreas de demanda.

HOGAN & REVELLE (1986) estenderam BACOP1 para uma situação na qual a primeira cobertura não é mandatária, mas apenas um objetivo, tal como em PLMC. Além disso uma segunda cobertura é também um objetivo, o que deu origem ao modelo BACOP2, definido matematicamente por:

$$\text{Maximize} \quad Z_1 = \sum_{j \in J} f_j x_j , \tag{32}$$

$$Z_2 = \sum_{j \in J} f_j u_j , \tag{33}$$

$$\text{sujeito a} \quad u_j + x_j \leq \sum_{i \in I} a_{ij} y_i , j \in J, \tag{34}$$

$$u_j \leq x_j , j \in J, \tag{35}$$

$$\sum_{i \in I} y_i = p, \tag{36}$$

$$y_i \in \{0,1\}, i \in I, \tag{37}$$

$$x_j, u_j \in \{0,1\}, j \in J. \tag{38}$$

O leitor deverá observar que, diferentemente de BACOP1**,** em BACOP2 o valor de *p* representa na realidade uma restrição de investimento (sendo portanto determinado apenas em função dos recursos disponíveis), pois a primeira cobertura para cada área de demanda não é mandatária. As restrições (34) asseguram que uma área de demanda $j \in J$ possui cobertura inicial se existe pelo menos um servidor, e dupla cobertura se existem pelo menos dois servidores a menos de *S* da mesma. As restrições (35) explicitam que para uma área de demanda ter dupla cobertura é necessário que ela tenha a primeira cobertura.

BACOP2 é um modelo com dois objetivos e como tal não tem solução ótima única. Nesses modelos há uma compensação (*trade-off*) entre os dois objetivos; não é possível maximizar simultaneamente tanto a primeira quanto a segunda cobertura. Em outras palavras, um aumento da população com cobertura inicial acarreta uma diminuição da população com cobertura adicional; por outro lado, um aumento da população com dupla cobertura acarreta um aumento da população sem nenhuma cobertura.

4.3 Modelos de Cobertura Probabilísticos

Embora para sistemas congestionados os modelos com cobertura adicional representem um avanço em relação ao Problema de Localização de Máxima Cobertura (PLMC), um desenvolvimento natural da teoria relacionada à localização de instalações de emergência foram os modelos de cobertura probabilísticos. Os modelos probabilísticos mostrados a seguir levam em conta a aleatoriedade na disponibilidade dos servidores, que é o fator mais importante em modelos de cobertura.

Estudaremos inicialmente o *Problema de Localização de Máxima Disponibilidade (PLMD)*, que é derivado do PLMC, seu equivalente determinístico. O PLMD incorpora hipóteses simplificadoras, o que, pela definição de *restrições de oportunidade* (*chance constraints*), permite definir uma formulação determinística equivalente para o mesmo. Este problema busca localizar *p* servidores tal que o máximo número de chamadas a um serviço de emergência tenha um servidor disponível a menos da distância crítica *S* com confiabilidade α. Ou seja, deseja-se que uma chamada a um serviço de emergência encontre um servidor disponível a menos da distância *S* com probabilidade $P \geq \alpha$, sendo α um valor predefinido.

O PLMD foi introduzido por REVELLE & HOGAN (1989a), sendo apresentado em duas versões, ambas incorporando hipóteses simplificadoras, que diferem no entanto na maneira pela qual a fração do tempo em que os servidores estão ocupados é calculada. A primeira versão, PLMDI, parte da hipótese que todos os servidores têm a mesma taxa de ocupação, igual à taxa de ocupação média do sistema, ρ. Já na segunda versão, PLMDII, calcula-se de forma diferenciada a fração de tempo em que os servidores estão ocupados, com valores específicos para diferentes setores da região geográfica em consideração.

Note que a definição de taxas de ocupação específicas para diferentes regiões geográficas em PLMDII não é equivalente a definir taxas de ocupação por servidor, nem permite o cálculo de tais taxas, o que exigiria que o modelo hipercubo fizesse parte do método de solução (veja seções 4.4 e 5). Assim, embora a utilização de taxas de ocupação específicas para cada região represente um avanço em relação à hipótese simplificadora de PLMDI, PLMDII também incorpora hipóteses simplificadoras. No presente trabalho nos limitaremos no entanto a analisar PLMDI.

Em PLMDI, ρ, a fração do tempo em que os servidores estão ocupados, é estimada por uma fórmula desenvolvida por DASKIN (1982). Sejam ϕ_j o número de chamadas por dia originadas na área de demanda $j \in J$ e $\bar{t}$ a duração média do atendimento a uma chamada, medida em horas. Então ρ pode ser calculada por

$$\rho = \frac{\bar{t}.\sum_{j\in J}\phi_j}{24.\sum_{i\in I} y_i} = \frac{\bar{t}.\sum_{j\in J}\phi_j}{24.p},$$

isto é, a fração ocupada de cada servidor é calculada como sendo a divisão do número médio de horas de serviço necessárias por dia no sistema pelo número de horas diárias disponíveis, levando-se em conta que serão localizados p servidores.

A restrição de que pelo menos um servidor deve estar disponível a menos de S de uma área de demanda $j \in J$, com probabilidade maior ou igual a α, pode ser escrita da seguinte forma:

P [um ou mais servidores disponíveis a menos de S] $\geq \alpha \equiv$

$\equiv$ (1 - P [nenhum servidor disponível a menos de S]) $\geq \alpha =$

$= 1 - \rho^{\sum_{i\in I} a_{ij} y_i} \geq \alpha$,

onde $\sum_{i\in I} a_{ij} y_i$ é o número de servidores disponíveis a menos de S da área de demanda $j \in J$. Ou, tomando logaritmos,

$$\sum_{i\in I} a_{ij} y_i \geq b, \quad \text{onde} \quad b = \left\lceil \frac{\log(1-\alpha)}{\log \rho} \right\rceil, \quad \lceil n \rceil$$

denotando o menor inteiro maior ou igual a n.

Da expressão linear equivalente obtida para a restrição probabilística é possível notar que cada área de demanda $j \in J$ exige a presença de pelo menos b servidores disponíveis a menos da distância crítica S para que seja assegurada uma cobertura da mesma com confiabilidade α. Portanto, para maximizar o número de chamadas com um serviço de confiabilidade α , é necessário maximizar o número de chamadas com pelo menos b servidores disponíveis a menos da distância S.

Seja a variável $x_{jk} \in \{0,1\}$ tal que $x_{jk} = 1$ se a área de demanda j tem pelo menos k servidores a menos de S, $x_{jk} = 0$ caso contrário. A expressão $\sum_{k=1}^{n} x_{jk}$ representa o número de vezes que a área de demanda $j \in J$ é coberta a menos de S. É finalmente possível escrever a seguinte expressão, que conta o número de coberturas disponíveis para cada área de demanda $j \in J$ com confiabilidade α: $\sum_{k=1}^{b} x_{jk} \leq \sum_{i\in I} a_{ij} y_i, j \in J$. Para maximizar o número de chamadas cobertas com confiabilidade α devemos maximizar $\sum_{j\in J} \phi_j x_{jb}$.

A formulação matemática de PLMDI pode ser finalmente escrita da seguinte forma:

$$\text{Maximize } Z = \sum_{j \in J} \phi_j x_{jb}, \quad (39)$$

$$\text{sujeito a } \sum_{k=1}^{b} x_{jk} \leq \sum_{i \in I} a_{ij} y_i, j \in J, \quad (40)$$

$$x_{jk} \leq x_{j(k-1)}, j \in J, k = 2,...,b, \quad (41)$$

$$\sum_{i \in I} y_i = p, \quad (42)$$

$$y_i, x_{jk} \in \{0,1\}, i \in I, j \in J, k=2,..b. \quad (43)$$

As restrições (40) asseguram que uma área de demanda $j \in J$ tem cobertura com confiabilidade α se existirem pelo menos b servidores a menos de S da mesma. Já as restrições (41) expressam o fato que para uma área de demanda ser coberta por k servidores ela tem que ser coberta por pelo menos (k-1) servidores, para $2 \leq k \leq b$.

DASKIN (1983) definiu o *Problema de Máxima Cobertura Esperada* (*PMCE*), que, tal como PLMD, estende PLMC para o caso probabilístico, mas com uma função objetivo diferente. No PMCE o objetivo é maximizar o aumento na cobertura esperada para todos os nós que resulta da disponibilidade de um servidor k adicional capaz de cobrir cada nó da rede. DASKIN parte da hipótese de independência entre servidores e, tal como em PLMDI, que todos os servidores têm a mesma taxa de ocupação ρ. Em seu modelo, no entanto, DASKIN permite que mais de um servidor seja localizado em um mesmo nó.

Utilizando a mesma notação de PLMD, a probabilidade que um nó j seja coberto por pelo menos um servidor, dado que m servidores cobrem este nó a menos da distância crítica, é dada por

P [um ou mais servidores disponíveis a menos de S] =
= (1 - P [nenhum servidor disponível a menos de S])
= $1 - \rho^m$.

Seja $H_{j,m}$ uma variável aleatória igual à demanda do nó j coberta por um servidor disponível, dado que m servidores cobrem este nó. Temos que $H_{j,m} = f_j$ com probabilidade $(1 - \rho^m)$, $H_{j,m} = 0$ com probabilidade ρ^m. O valor esperado de $H_{j,m}$ é dado por

$$E(H_{j,m}) = f_j (1 - \rho^m) \; \forall \, j,m.$$

O aumento na cobertura esperada no nó j ao se aumentar o número de servidores que o cobrem de $(m - 1)$ para m é dado por

$$\Delta E(H_{j,m}) = E(H_{j,m}) - E(H_{j,m-1}) = f_j \, \rho^{m-1} (1 - \rho), \; m = 1, 2, ..., p.$$

Utilizando as definições acima o PMCE pode ser formulado como um problema de programação inteira:

$$\text{Maximize } Z = \sum_{j \in J} \sum_{k=1}^{p} (1 - \rho) \rho^{k-1} f_j x_{jk}, \quad (44)$$

$$\text{sujeito a } \sum_{k=1}^{p} x_{jk} \leq \sum_{i \in I} a_{ij} y_i, j \in J, \quad (45)$$

$$\sum_{i \in I} y_i \leq p, \quad (46)$$

$$y_i = 0,1,..., p, i \in I, \quad (47)$$

$$x_{jk} \in \{0,1\}, \forall j,k. \quad (48)$$

A formulação acima é semelhante à de PLMDI, exceto evidentemente pela função objetivo e pelo fato que até p servidores podem ser localizados em um dado nó da rede. Observe-se também a diferença entre as restrições (40) e (45); por outro lado, como a função objetivo é côncava em k para cada j (ver DASKIN, 1983), não é necessário explicitar, em PMCE, restrições de precedência tal como as restrições (41) de PLMDI.

A localização de serviços de emergência tem gerado um número significativo de artigos em modelos de localização probabilística. DASKIN *et al.* (1988) fazem uma revisão bibliográfica enfocando a disponibilidade de servidores em problemas com restrições de cobertura. REVELLE & HOGAN (1989b) desenvolveram dois modelos adicionais que estudam a disponibilidade de servidores neste mesmo contexto, o

problema dos p centros com α-confiabilidade e o problema de localização de máxima confiabilidade. O problema dos p-centros com α-confiabilidade, que estende o problema dos p-centros, busca localizar p facilidades de forma a minimizar o tempo máximo até um dado serviço estar disponível com confiabilidade α, isto é, dadas p facilidades a serem localizadas, devemos fazê-lo de modo que o tempo máximo decorrido até o serviço poder ser utilizado (ele está disponível durante uma fração α do tempo) seja o menor possível. Já o problema de localização de máxima confiabilidade busca localizar p facilidades de modo que a probabilidade de um dado serviço estar disponível dentro de um tempo (ou distância) pré-especificado seja a maior possível. DASKIN & HAGHANI (1984) consideram casos em que múltiplos veículos são despachados para atender uma dada emergência.

Conforme já discutido, o uso de hipóteses simplificadoras, tal como em PLMD e PCME, permite a formulação de problemas de localização probabilísticos através da programação matemática. A ausência de tais hipóteses, que aproxima a modelagem de situações encontradas no mundo real, conduz ao tratamento desses problemas através de modelos de filas espacialmente distribuídas.

4.4 Modelos que Utilizam Teoria das Filas

O modelo hipercubo de LARSON (1974), examinado nas seções 2 e 3 deste artigo, possibilita a utilização da teoria das filas em modelos de localização probabilísticos. O modelo hipercubo é um modelo descritivo, que permite o cálculo de medidas de desempenho que caracterizam um dado sistema com base em filas espacialmente distribuídas. Aplicado de forma isolada não assegura a solução direta de problemas de localização, sendo no máximo uma ferramenta que pode ser aplicada ao estudo de diversos cenários. Este modelo tem sido no entanto embutido por alguns autores em métodos heurísticos desenvolvidos para resolver problemas de localização probabilísticos.

BATTA *et al.* (1989) se ocupam das seguintes três hipóteses do PMCE de DASKIN: (i) os servidores operam de forma independente; (ii) todos os servidores têm a mesma taxa de ocupação ρ; (iii) as taxas de ocupação não variam com a localização dos servidores. Os autores partem do princípio que a cooperação entre servidores sempre acontece na prática, dada a necessidade de atendimento rápido a chamadas de emergência. Eles provam que tal cooperação invalida a hipótese de independência entre os servidores feita por DASKIN. Relaxando então as três hipóteses do PCME, e partindo do princípio que o sistema de filas está em equilíbrio e que as hipóteses básicas do modelo hipercubo são aplicáveis, BATTA *et al.* embutem este modelo em uma heurística de substituição de vértices, buscando determinar a localização dos servidores que maximize a cobertura esperada.

Os autores partem de uma localização inicial para os p servidores e resolvem o modelo hipercubo para esta configuração, calculando a cobertura esperada correspondente. É aplicado então o algoritmo de substituição de vértices, desenvolvido originalmente por TEITZ & BART (1968) para o problema das p-medianas, buscando achar configurações alternativas que melhorem o valor da cobertura esperada. Para cada configuração considerada é necessário resolver o modelo hipercubo, para que a cobertura esperada correspondente possa ser calculada.

A heurística de substituição de vértices prossegue até que nenhuma troca simples de localização de servidores (mudança de localização de um único servidor) produza uma melhora no valor da cobertura esperada. O uso do modelo hipercubo permite que sejam consideradas taxas de ocupação individualizadas para cada servidor (ao invés de uma taxa única ρ, como por exemplo no PCME de DASKIN), aumentando a precisão dos cálculos.

BERMAN *et al.* (1985) examinam o problema de localizar um único servidor em redes congestionadas considerando explicitamente o processo da chegada de chamadas por serviço.

Os autores definem dois modelos que são extensões do problema das *p*-medianas definido por HAKIMI (1964, 1965). No primeiro modelo, chamado de *problema das medianas com perda estocástica*, as chamadas que encontram o servidor ocupado são perdidas; no segundo, chamado de *problema das medianas com fila estocástica*, essas chamadas entram em uma fila que é atendida de acordo com a disciplina FIFO. O objetivo no primeiro problema é minimizar a soma ponderada do tempo médio de viagem e dos custos de rejeição; no segundo, o objetivo é minimizar a soma do tempo médio de viagem com o tempo médio na fila.

Não é nossa intenção neste artigo fazer uma revisão bibliográfica abrangente de modelos de localização que utilizam teoria das filas para sua formulação e solução, mas apenas exemplificar o uso do modelo hipercubo nesse contexto. Revisões bibliográficas sobre o tópico são feitas, de forma bastante competente, por SWERSEY (1994), e mais recentemente por OWEN & DASKIN (1998); o leitor interessado deve consultar essas referências para maiores detalhes.

5. Métodos de Solução para Problemas Probabilísticos

Problemas de localização probabilísticos são em geral resolvidos por métodos heurísticos. No caso dos modelos formulados através da programação matemática, a obtenção da solução ótima é em princípio possível, quer utilizando "pacotes" computacionais como pelo desenvolvimento de métodos especializados, mas na prática o uso de heurísticas prevalece. No caso dos modelos com base em teoria das filas a utilização de métodos exatos parece irrealista, pois qualquer formulação matemática que inclua o modelo hipercubo embutido dificilmente será matematicamente tratável. Neste caso a única possibilidade é o uso de métodos aproximados de solução.

Como os dois modelos probabilísticos inicialmente apresentados na seção 4, PLMD e PCME, podem ser ambos considerados versões probabilísticas do PLMC, faz sentido examinar inicialmente métodos de solução disponíveis para este problema determinístico. Métodos de solução inicialmente propostos para o PLMC incluem a relaxação de Programação Linear (PL) da formulação de programação 0-1 do problema (dada por (22)-(26) acima) e um método que combina heurística gulosa com substituição de vértices (veja CHURCH & REVELLE, 1974). GALVÃO & REVELLE (1996) desenvolveram uma heurística Lagrangeana para o problema; esses autores apresentam experiência computacional usando tanto dados da literatura como redes geradas aleatoriamente.

Métodos exatos incluem o algoritmo de DWYER & EVANS (1981), desenvolvido para o caso particular em que todos os pontos geradores de demanda têm o mesmo peso, e o algoritmo com base no dual de DOWNS & CAMM (1996). Estes últimos autores apresentam uma avaliação computacional extensiva de seu método, tanto em termos de variedade de aplicações como de tamanho do problema. Finalmente, GALVÃO *et al.* (2000) comparam heurísticas que têm por base as relaxações Lagrangeana e *surrogate* do problema. Ambas as heurísticas produzem soluções de boa qualidade em tempos computacionais semelhantes. Os autores demonstram que quando os multiplicadores iniciais são escolhidos de forma apropriada, a heurística *surrogate* perde a vantagem computacional relatada por LORENA & LOPES (1994) e LORENA & NARCISO (1996) em outras aplicações do método.

5.1 Os Modelos de Cobertura Probabilísticos PMCE e PLMD

O PMCE, definido por DASKIN (1983), foi o primeiro modelo probabilístico derivado do PLMC a aparecer na literatura. DASKIN desenvolveu uma heurística de substituição de vértices para resolver esse problema. O algoritmo é iniciado sob a hipótese que todos os servidores estão ocupados praticamente durante o tempo todo (ρ muito próximo de 1). Nessas condições o

autor argumenta que todos os servidores devem ser localizados no nó que cobre a maior parte da demanda. Sua heurística então usa substituição de vértices (trocando sempre um único vértice por outro em cada tentativa) para calcular taxas de ocupação para as quais as localizações (de alguns) dos servidores devem mudar. O algoritmo de substituição de vértices prossegue, achando a melhor configuração da localização dos servidores para vários valores de ρ ($0 \le \rho \le 1$).

DASKIN testou seu algoritmo para a rede de 55 nós definida por SWAIN (1971). Embora não garanta a otimalidade das soluções encontradas, o algoritmo pode detectar subotimalidades, em cujo caso análises pós-heurística muitas vezes permitem obter soluções de melhor qualidade. Outra observação importante é que para $\rho = 0$ o problema se reduz ao PLMC (determinístico). A solução obtida por DASKIN (1983) para este caso, para a instância testada em seu artigo ($n = 55$, $p = 3$, $S = 15$), é ótima para o MCLP, conforme constatamos ao rodar o código desenvolvido por GALVÃO & REVELLE (1996).

BATTA *et al.* (1989) "revisitaram" o PCME. Conforme já observado na seção 4.4, os autores relaxaram as três hipóteses simplificadoras utilizadas por DASKIN e embutiram o modelo hipercubo em uma heurística de substituição de vértices, desenvolvendo um algoritmo para o cálculo da cobertura máxima esperada mais adaptado às condições encontradas em aplicações práticas. BATTA *et al.* também sugerem que uma maneira aproximada de relaxar a hipótese de independência entre servidores é utilizar os fatores de correção desenvolvidos por LARSON (1975) para o modelo hipercubo (veja a seção 3.2). Estes fatores de correção, aplicados à função objetivo do PCME definido por DASKIN, levam ao modelo PCME ajustado, que chamaremos de PCMEA.

De forma a melhorar os resultados obtidos por DASKIN, BATTA *et al.* resolvem de forma exata (por programação inteira) os modelos PCME e PCMEA, para valores específicos de ρ, em um procedimento que eles chamam de análise pós-PI (pós-programação inteira). Finalmente os autores comparam, para a rede de 55 vértices utilizada por DASKIN, os seguintes procedimentos: (i) heurística e análise pós–heurística de DASKIN, seguida da análise pós-PI proposta por eles, aplicadas às funções objetivo definidas para PCME e PCMEA; (ii) substituição de vértices com o modelo hipercubo embutido, desenvolvido pelos autores.

As conclusões dos autores são no sentido que as discrepâncias entre as coberturas esperadas calculadas pelos três procedimentos (PCME, PCMEA e substituição de vértices) são devidas às hipóteses simplificadoras de PCME e PCMEA: PCME superestima as coberturas esperadas, com o valor do erro aumentando com ρ; PCMEA, em menor grau, superestima as coberturas esperadas para valores pequenos de ρ e as subestima para valores maiores da taxa de ocupação dos servidores. As localizações dos servidores obtidas pelos três procedimentos são no entanto fisicamente bastante próximas.

A questão do cálculo da cobertura esperada, importante por exemplo na aplicação de modelos de probabilísticos à localização de serviços médicos de emergência, foi retomada por SAYDAM *et al.* (1994). Os autores utilizaram os modelos propostos por DASKIN (1983) e BATTA *et al.* (1989), entre outros, a diversos cenários simulados de serviços médicos de emergência. Eles concluem que nenhum dos modelos de cobertura estudados é consistentemente mais preciso que os demais. Os resultados obtidos por SAYDAM *et al.* apóiam no entanto a recomendação de BATTA *et al.*, no sentido que um modelo com base no hipercubo seja utilizado para análises de pós-otimização.

BATTA *et al.* resolvem o modelo hipercubo, nas diversas iterações de sua heurística de substituição de vértices, pelo método aproximado de Larson (veja seção 3.2). Uma alternativa possível, se bem que computacionalmente mais cara, é resolver o modelo hipercubo de forma exata em cada iteração, observando o possível impacto desta estratégia no valor da cobertura máxima esperada.

REVELLE & HOGAN (1989a) resolveram o PLMD, nas versões PLMDI e PLMDII, para dados do sistema de combate a incêndios da cidade de Baltimore, nos Estados Unidos. Esses dados consistem de 207 áreas de planejamento, com a freqüência média de chamadas por unidade de tempo conhecida para cada área. Eles analisaram a localização de ambulâncias nesse sistema, com uma ou mais ambulâncias podendo ser localizadas em cada uma das 31 estações do corpo de bombeiros existentes na região em estudo. Para resolver o problema os autores utilizaram o "pacote" de programação matemática MPSX em um microcomputador IBM AT/370. Infelizmente, no entanto, não são fornecidos detalhes dos resultados computacionais obtidos, o que impede uma avaliação mais acurada da metodologia utilizada.

5.2 Outros Modelos Probabilísticos

Para resolver os problemas de mediana com perda estocástica e com fila estocástica, BERMAN *et al.* (1985) modelam ambos os sistemas como uma fila *M/G/1* operando em equilíbrio, com o tamanho máximo da fila igual a zero e infinito, respectivamente. Os autores demonstram que a solução do problema com perda estocástica se reduz ao problema padrão de *1*-mediana de HAKIMI, enquanto que o problema com fila estocástica tem função objetivo não linear, o que pode levar à localização ótima em qualquer ponto da rede, nos nós ou no interior dos arcos da mesma. Os autores desenvolveram um método exato que acha a localização ótima da mediana com fila estocástica em um número finito de iterações.

A tentativa de generalizar a análise do *problema das medianas com fila estocástica* para o caso de p servidores esbarra em dificuldades teóricas. BERMAN *et al.* (1987) estudaram esse problema, tendo desenvolvido duas heurísticas para localizar os p servidores. A primeira heurística parte de um conjunto inicial de p pontos e usa o modelo hipercubo para fornecer a cada servidor informação sobre a probabilidade de ser despachado para atender cada ponto gerador de demanda. Esta informação é então utilizada na solução de problemas de *1*-mediana, utilizados para melhorar a localização de cada servidor. A segunda heurística é semelhante à primeira, exceto que problemas de medianas com fila estocástica são resolvidos ao invés de problemas de *1*-mediana. Ambas as heurísticas utilizam conceitos semelhantes aos usados por MARANZANA (1964) em sua heurística desenvolvida para o problema das p-medianas, e ambas produzem soluções de boa qualidade.

Tanto o problema dos p-centros com α-confiabilidade como o problema de localização de máxima confiabilidade, definidos por REVELLE & HOGAN (1989b), podem ser formulados e resolvidos usando-se um procedimento desenvolvido pelos autores para estimar as taxas de ocupação dos servidores em cada área geradora de demanda. Segundo os mesmos, ambos envolvem a solução de uma seqüência de problemas de localização para a cobertura de conjuntos (PLCC's) probabilísticos. O PLCC probabilístico foi originalmente definido por CHAPMAN & WHITE (1974).

Finalmente REVELLE & HOGAN (1989b), ao analisar o modelo PLMDII, observam que taxas de ocupação individualizadas, para servidores localizados em áreas específicas, podem ser consideradas em um modelo do tipo PLMD através da utilização do modelo hipercubo. Pesquisa nesse sentido está sendo conduzida atualmente no âmbito de tese de doutorado orientada por um dos autores. O método utilizado é uma heurística de substituição de vértices que embute o modelo hipercubo, na tentativa de localizar os servidores de modo a maximizar o nível de confiabilidade α definido para o modelo.

6. Conclusões

O uso do modelo hipercubo em métodos de solução para problemas de localização probabilísticos, embora relatado na literatura especializada, é ainda bastante incipiente e

limitado. É de bastante importância, no entanto, em situações nas quais a aleatoriedade na disponibilidade dos servidores é um fator importante a ser considerado. Em algumas circunstâncias esta aleatoriedade pode ser melhor representada pela modelagem de filas espacialmente distribuídas.

O presente artigo trata da integração do modelo hipercubo à modelagem probabilística de problemas de localização. Foi visto que neste caso a utilização de métodos exatos de solução parece irrealista, pois qualquer formulação matemática que inclua o modelo hipercubo embutido dificilmente será matematicamente tratável. A única possibilidade é então o uso de métodos de solução aproximados.

As poucas aplicações relatadas na literatura sobre a utilização do modelo hipercubo neste contexto referem-se a métodos de substituição de vértices. Nestas aplicações o modelo hipercubo faz o papel de "sub-rotina", responsável pelo cálculo da função objetivo (utilizando as medidas de desempenho calculadas pelo modelo) para cada configuração de localização dos servidores.

Outras possibilidades nos parecem no entanto evidentes, por exemplo a utilização do hipercubo em metaheurísticas tais como *simulated annealing* e busca tabu. Nosso objetivo foi estimular a discussão em torno do assunto através deste trabalho, que tem caráter tutorial quanto ao modelo hipercubo, e caráter de revisão bibliográfica no que se refere aos modelos de localização probabilísticos apresentados.

Agradecimentos

Os autores são gratos aos três revisores anônimos pelos úteis comentários e sugestões.

Referências Bibliográficas

ALBINO, J.C.C.: *Quantificação e locação de unidades móveis de atendimento de emergência e interrupções em redes de distribuição de energia elétrica: aplicação do modelo hipercubo*. Tese de Mestrado, Depto. de Engenharia de Produção e Sistemas, Universidade Federal de Santa Catarina, 1994.

ATKINSON, K.E.: *An Introduction to Numerical Analysis*, John Wiley & Sons, 1978.

BATTA, R.; DOLAN, J.M. & KRISHNAMURTHY, N.N.: "The maximal expected covering location problem: Revisited". *Transportation Science*, **23**, 277-287, 1989.

BERMAN, O.; LARSON, R.C. & CHIU, S.S.: "Optimal server location on a network operating as an *M/G/1* queue". *Operations Research*, **33**, 746-771, 1985.

BERMAN, O.; LARSON, R.C. & PARKAN, C.: "The stochastic queue *P*-median problem". *Transportation Science*, **21**, 207-216, 1987.

BRANDEAU, M.L. & LARSON, R.C.: "Extending and applying the hipercube model to deploy ambulances in Boston". **In**: SWERSEY, A. & IGNALL, E. (ed.). *Delivery of Urban Services – TIMS Studies in the Management Sciences*, *Vol. 22*, Elsevier Science, B.V., 121-153, 1986.

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

CHAPMAN, S. & WHITE, J.: "Probabilistic formulations of emergency service facilities location problems". *ORSA/TIMS paper*, San Juan, Puerto Rico, 1974.

CHURCH, R. & REVELLE, C.S.: "The maximal covering location problem". *Papers of the Regional Science Association*, **32**, 101-118, 1974.

DASKIN, M.S.: "Application of an expected covering model to emergency medical service system design". *Decision Sciences*, **13**, 416-439, 1982.

DASKIN, M.S.: "A maximum expected covering location model: Formulation, properties and and heuristic solution". *Transportation Science*, **17**, 48-70, 1983.

DASKIN, M.S. & HAGHANI, A.: "Multiple vehicle routing and dispatching to an emergency scene". *Environment and Planning A*, **16**, 1349-1359, 1984.

DASKIN, M.S.; HOGAN, K. & REVELLE, C.S.: "Integration of multiple, excess, backup and covering models". *Environment and Planning B*, **15**, 15-35, 1988.

DOWNS, B.T. & CAMM, J.D.: "An exact algorithm for the maximal covering location problem". *Naval Research Logistics Quarterly*, **43**, 435-461, 1996.

DWYER, F.R. & EVANS, J.R.: "A branch and bound algorithm for the list selection problem in direct mail advertising". *Management Science*, **27**, 658-667, 1981.

GALVÃO, R.D. & REVELLE, C.S.: "A Lagrangean heuristic for the maximal covering location problem". *European Journal of Operational Research*, **88**, 114-123, 1996.

GALVÃO, R.D.; ESPEJO, L.G.A. & BOFFEY, B.: "A comparison of Lagrangean and surrogate relaxations for the maximal covering location problem". *European Journal of Operational Research*, **124**, 159-171, 2000.

GONÇALVES, M.B.; NOVAES, A.G.N. & ALBINO, J.C.C.: "Modelos para localização de serviços emergenciais em rodovias". *Anais do XXVI Simpósio Brasileiro de Pesquisa Operacional*, **1**, 591-596, 1994.

GONÇALVES, M.B.; NOVAES, A.G.N. & SCHMITZ, R.: "Um modelo de otimização para localizar unidades de serviços emergenciais em rodovias". *Anais do IX ANPET – Congresso de Pesquisa e Ensino em Transportes*, 962-972, 1995.

GROSS, D. & HARRIS, C.M.: *Fundamentals of Queueing Theory*. John Wiley & Sons, New York, 1974.

HAKIMI, S.L.: "Optimum locations of switching centers and the absolute centers and medians of a graph". *Operations Research*, **12**, 450-459, 1964.

HAKIMI, S.L.: "Optimum distribution of switching centers in a communication network and some related graph theoretic problems". *Operations Research*, **13**, 462-475, 1965.

HALPERN, J.: "Accuracy of estimates for the performance criteria in certain emergency service queueing systems". *Transportation Science*, **11**, 223-242, 1977.

HOGAN, K. & REVELLE, C.S.: "Concepts and applications of backup coverage". *Management Science*, **32**, 1434-1444, 1986.

LARSON, R.C.: "A hypercube queuing model for facility location and redistricting in urban emergency services". *Computers and Operations Research*, **1**, 67-95, 1974.

LARSON, R.C.: "Approximating the performance of urban emergency service systems". *Operations Research*, **23**, 845-868, 1975.

LARSON, R.C. & ODONI, A.R.: *Urban Operations Research*, Prentice Hall, Inc., N.J, 1981.

LORENA, L.A.N. & LOPES, F.B.: "A surrogate heuristic for set covering problems". *European Journal of Operational Research*, **79**, 138-150, 1994.

LORENA, L.A.N. & NARCISO, M.G.: "Relaxation heuristics for a generalized assignment problem". *European Journal of Operational Research*, **91**, 600-610, 1996.

MARANZANA, F.E.: "On the location of supply points to minimize transport costs". *Operational Research Quarterly*, **15**, 261-270, 1964.

MENDONÇA, F.C.: *Aplicação do modelo hipercubo, baseado em teoria de filas, para análise de um sistema médico-emergencial em rodovia*. Tese de Mestrado, Depto. de Engenharia de Produção, Universidade Federal de São Carlos, 1999.

MENDONÇA, F.C. & MORABITO, R.: "Aplicação do modelo hipercubo para análise de um sistema médico-emergencial em rodovia". *Gestão & Produção*, **7**, 173-91, 2000.

OWEN, S.H. & DASKIN, M.S.: "Strategic facility location: A review". *European Journal of Operational Research*, **111**, 423-447, 1998.

REVELLE, C.S. & HOGAN, K.: "The maximum availability location problem". *Transportation Science*, **23**, 192-200, 1989a.

REVELLE, C.S. & HOGAN, K.: "The maximum reliability location problem and α-reliable p-center problem: Derivatives of the probabilistic location set covering problem". *Annals of Operations Research*, **18**, 155-174, 1989b.

SACKS, S.R. & GRIEF, S.: "Orlando Police Department uses OR/MS methodology, new software to design patrol districts". *OR/MS Today*, 30-32, Feb. 1994.

SAYDAM, C.; REPEDE, J.F. & BURWELL, T.: "Accurate estimation of expected coverage – a comparative study". *Socio-Economic Planning Sciences*, **28**, 113-120, 1994.

SWAIN, R.: A decomposition algorithm for a class of facility location problems. Ph.D. Thesis, Cornell University, Ithaca, NY, 1971.

SWERSEY, A.J.: "The deployment of police, fire and emergency medical units". **In:** POLLOCK, S.M. *et al.* (ed.). *Handbooks in OR & MS, Vol. 6*, Elsevier Science B.V., 151-200, 1994.

TAKEDA, R.: *Tese de Doutorado*, Engenharia de Transportes, Escola de Engenharia de São Carlos, USP. Em preparação, previsão de defesa para segundo semestre de 2000.

TEITZ, M.B. & BART, P.: "Heuristic methods for estimating the generalized vertex median of a weighted graph". *Operations Research*, **16**, 955-961, 1968.

THE USE OF THE HYPERCUBE MODEL IN THE SOLUTION OF PROBABILISTIC LOCATION PROBLEMS

Abstract

The hypercube model is revisited regarding its use in solution methods for probabilistic location problems. This use of the model is relevant in situations in which the randomness in the availability of servers is an important factor to be considered; in some circumstances this randomness can be represented by spatially distributed queues. The model is presented through an illustrative example, for which the equilibrium equations are derived; some measures of performance are also defined. This is followed by the description of an exact and an approximate method for the calculation of these measures. Several probabilistic location models are then studied, which is followed by the analysis of solution methods for these models, with special emphasis given to methods that embed the hypercube model. Although incipient at present, the use of the hypercube model in probabilistic location problems has good potential, for example if embedded into metaheuristics such as simulated annealing and tabu search.

Key words: location, hypercube model, probabilistic models, queues.