**UNIVERSIDADE FEDERAL DE SÃO CARLOS**
**CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA**
**DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO**

# OTIMIZAÇÃO DA CONFIGURAÇÃO E OPERAÇÃO DE SISTEMAS MÉDICO EMERGENCIAIS EM RODOVIAS UTILIZANDO O MODELO HIPERCUBO

Ana Paula Iannoni

Tese apresentada à Universidade

Federal de São Carlos, como parte dos requisitos para obtenção do título de Doutora em Engenharia de Produção

ORIENTADOR : Prof. Dr. Reinaldo Morabito

São Carlos – S.P

Março/2005

**Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária/UFSCar**

I11oc

Iannoni, Ana Paula.
Otimização da configuração e operação de sistemas médicos emergenciais em rodovias utilizando o modelo hipercubo / Ana Paula Iannoni. -- São Carlos : UFSCar, 2005.
229 p.

Tese (Doutorado) -- Universidade Federal de São Carlos, 2005.

1. Logística empresarial. 2. Rodovias. 3. Atendimento médico - emergencial. 4. Modelo hipercubo. I. Título.

CDD: 658.5 (20ª)

# Agradecimentos

A Deus.

Ao meu orientador Reinaldo Morabito, pela orientação, paciência, incentivo, confiança e dedicação de sempre, e pela amizade devotada durante todo estes anos em que trabalhamos juntos. A sua maneira sábia e cautelosa de encarar os problemas trouxe-se me ânimo nos momentos mais difíceis.

Ao Prof. Cem Saydam por me receber como pesquisadora visitante para conduzir estudos sob sua supervisão. Em especial, agradeço-o muito pelo suporte, incentivo, troca de conhecimentos e principalmente amizade

Ao CNPQ pela bolsa de estudos concedida no país, sem a qual este trabalho não teria sido realizado.

A CAPES pela bolsa de doutorado sanduíche, que viabilizou a oportunidade de trabalhar e trocar experiências com especialistas na minha área de interesse no exterior.

À concessionária *Centrovias* pela valiosa e atenciosa colaboração durante o desenvolvimento esta pesquisa viabilizando o estudo de caso por meio de visitas e coleta de dados em seus centros administrativo e operacional.

Aos meus amigos de pós-graduação do Departamento de Engenharia de Produção da UFSCar, pelo carinho e amizade. E aos amigos da UNCC, em especial Hari e Beth.

Gostaria também de expressar meus agradecimentos à Doris por toda amizade e carinho devotados.

Ao meu querido irmão João, pelo amor fraterno e paciência de cada dia e as minhas queridas amigas Jana e Cintia, por dividirem comigo momentos de alegria e estarem comigo nos momentos mais difíceis da minha vida.

E de forma muito especial, aos meus pais, João e Adair (*in Memorian*), pelo amor infinito, carinho e incentivo na realização dos meus sonhos. A saudade é grande, mas a meta é que de onde estejam tenham orgulho de seus filhos.

## Sumário

## Lista de figuras

## Lista de Tabelas

## Lista de principais símbolos e siglas:

$b_j$ estado do servidor $j$;

$B$ representação vetorial de um estado do sistema (i.e., $B = \{b_1, b_2, .., b_N\}$ );

$B_j$ estado em que somente o servidor $j$ está disponível no sistema;

$c$ índice utilizado para indicar solução;

CCO central de controle de operações da *Centrovias*;

$d_j$ corresponde a distância entre dois servidores adjacentes $j$ e $j+1$;

$\Delta$ valor de incremento para calcular proporção de distância entre servidores;

EMS *Emergency Medical Systems;*

$E_{ji}$ conjunto de estados nos quais o servidor $j$ é o primeiro servidor disponível na lista de despacho do átomo $i$;

$E_{(j,k)i}$ conjunto de estados nos quais os servidores $j$ e $k$ são os dois primeiros servidores disponíveis da lista de despacho do átomo $i$;

$f_{ji}$ fração de despachos no sistema que são atendidas pelo servidor $j$ no átomo $i$;

$f_{ji}^{[nq]}$ fração de despachos em que o servidor $j$ é enviado ao átomo $i$ que não implicam em tempo de espera na fila;

$f_{ji}^{[q]}$ fração de despachos em que o servidor $j$ é enviado ao átomo $i$ que implicam em tempo de espera na fila;

$f_I$ fração do total de despachos que foram realizados como *backup*;

$f_{Ij}$ fração de atendimentos realizados pelo servidor $j$ como *backup*;

$f_{Ii}^{[a]}$ fração das chamadas no átomo $i$ que foi atendida por servidores *backup*;

$f_{ji}^{[1]}$ fração total de despachos para atendimento de chamadas do tipo 1, nos quais o servidor $j$ é enviado ao átomo $i$;

$f_{(j,k)i}^{[2]}$ fração total de despachos para atendimento de chamadas tipo 2, nos quais os servidores $j$ e $k$ são enviados simultaneamente ao átomo $i$;

$f_{ji}^{[2]}$ fração total de despachos para atendimento de chamadas tipo 2, nos quais somente servidor $j$ é enviado ao átomo $i$;

$f'^{[1]}_{ji}$ freqüência de todos os despachos do sistema que envia o servidor $j$ ao átomo $i$, para atender uma chamada do tipo 1;

$f'^{[2]}_{(j,k)i}$ freqüência de todos os despachos do sistema que envia o servidor $j$ e $k$ ao átomo $i$, para atender uma chamada do tipo 2;

$f'^{[2]}_{ji}$ freqüência de todos os despachos do sistema que envia o servidor $j$ ao átomo $i$, para atender uma chamada do tipo 2;

$f'^{[m]}_{ji}$ freqüência de todos os despachos do sistema que envia o servidor $j$ para atender uma chamada to tipo $m$ no átomo $i$;

$f'^{[m]}_{(j,k)i}$ freqüência de todos os despachos do sistema que envia o servidor $j$ e $k$ para atender uma chamada to tipo $m$ no átomo $i$;

$f_{ij}^{[v]}$ corresponde a fração de todas as chamadas em que o servidor $j$ é enviado ao átomo $i$;

$F_{ji}$ conjunto de estados em que somente o servidor $j$ pode atender uma chamada em $i$;

$F_{(j,k)i}$ conjunto de estados em que os servidores $j$ e $k$ são os únicos servidores disponíveis na lista de preferência de despacho do átomo $i$;

GA algoritmo genético;

$G$ número de gerações;

$G_{ji}$ conjunto de estados nos quais o servidor preferencial $j$ do átomo $i$ está disponível;

$h_i$ demanda no átomo $i$;

$i$ índice em geral utilizado para identificar átomos;

$I$ conjuntos dos nós de demanda;

$j, k, l$ índices em geral utilizados para identificar servidores;

$K_i$ conjunto de servidores que oferecem cobertura ao nó de demanda $i$;

$L'_{ji}$ conjunto de servidores que possuem menor prioridade que o servidor $j$ na lista de despacho do átomo $i$;

$L_i^{[v]}$ conjunto de servidores da lista de despacho do átomo $i$ tal que $t_{ji} < t_{ki}$;

$\lambda$ taxa total de chegada no sistema;

$\lambda_i$ taxa de chegada no átomo $i$;

$\lambda_i^0$ taxa de chegada inicial do átomo $i$;

$\lambda_i^{[1]}$ taxa de chegada de chamadas tipo 1 no átomo $i$;

$\lambda_i^{[2]}$ taxa de chegada de chamadas tipo 2 no átomo $i$;

$\lambda^{[1]}$ taxa total de chegada de chamadas tipo 1 no sistema;

$\lambda^{[2]}$ taxa total de chegada de chamadas tipo 2 no sistema;

$\lambda_i^{[1a]}$ taxa de chegada de chamadas tipo 1a no átomo $i$;

$\lambda_i^{[2a]}$ taxa de chegada de chamadas tipo 2a no átomo $i$;

$\lambda_i^{[2b]}$ taxa de chegada de chamadas tipo 2b no átomo $i$;

$\lambda_i^{[3]}$ taxa de chegada de chamadas tipo 3 no átomo $i$;

$\lambda_i^{[m]}$ taxa de chegada de chamadas tipo $m$ no átomo $i$;

$m$ índice em geral utilizado para identificar tipo de chamada;

$\mu$ taxa total de atendimento no sistema;

$\mu_j$ taxa de atendimento do servidor $j$;

$\mu^{[I]}$ taxa total de atendimento para chamadas tipo 1 e 2 no sistema;

$\mu^{[II]}$ taxa total de atendimento para chamadas tipo 1a no sistema;

$\mu_j^{[I]}$ taxa de atendimento do servidor $j$ para chamadas tipo 1 e 2;

$\mu_j^{[II]}$ taxa de atendimento do servidor *j* para chamadas tipo 1a em sua base;

$n$ índice em geral utilizado para denotar número de usuários no sistema;

$N$ número de servidores no sistema;

$N_A$ número de átomos no sistema;

$N_{Aj}$ conjunto dos átomos para os quais o servidor $j$ é o servidor primário;

$N_A^{[b]}$ conjunto de átomos com todos os servidores de sua lista de preferência ocupados no estado *B*;

$p[j]$ área onde está localizado o servidor *j*;

$pr_i^{[2]}$ fração de múltiplo despacho do total de atendimentos em cada átomo *i*;

$p_c$ probabilidade de *crossover*;

$p_m$ probabilidade de mutação;

$P_B$ probabilidade de equilíbrio do estado *B*;

$P_S$ probabilidade do sistema estar saturado;

$P_q$ probabilidade de formação de fila;

$P_p$ probabilidade de perda de qualquer chamada do sistema;

$P_p^{[1]}$ probabilidade de perda de chamadas tipo 1;

$P_p^{[2]}$ probabilidade de perda de chamadas tipo 2;

$P_{t>\bar{tv}}$ probabilidade de uma chamada no sistema ser atendida em um tempo superior a $\bar{t}_v$ (limite predeterminado);

$p\left(t_{ji} > \bar{t}_v\right)$ corresponde a fração de chamadas no átomo *i* que esperam mais que $\bar{t}_v$ min pela chegada do servidor $j$;

$P_p^{[1a]}$ probabilidade de perda de chamadas tipo 1a;

$P_p^{[2a]}$ probabilidade de perda de chamadas tipo 2a;

$P_p^{[2b]}$ probabilidade de perda de chamadas tipo 2b;

$P_p^{[3]}$ probabilidade de perda de chamadas tipo 3;

$P_{\bar{tv}>10}(c)$ fração de chamadas que são atendidas em tempo superior à 10 minutos para a solução *c*;

$P_{op}$ tamanho da população;

$q_{jr}$ probabilidade do servidor $j$ estar no átomo $r$;

*r, s* índices em geral utilizados para denotar átomos;

$\rho$ carga média de trabalho do sistema;

$\rho_j$ carga de trabalho do servidor *j*;

$\rho_j^{[I]}$ carga de trabalho do servidor *j* atendendo uma chamada do tipo 1 ou tipo 2;

$\rho_j^{[II]}$ carga de trabalho do servidor *j* atendendo uma chamada tipo 1a;

$\bar{\rho}$ média das cargas de trabalho no solução *c;*

SAE sistemas de atendimento emergencial;

SAU serviço de atendimento ao usuário;

SAMU S*ervice d'Aide Médicale Urgente;*

$S_n$ estado em que há $n$ usuários no sistema, sendo que ($n$-N) estão em espera na fila;

$\sigma_\rho$ desvio padrão das cargas de trabalho entre os servidores do sistema;

$\sigma_\rho(c)$ desvio padrão das cargas de trabalho da solução $c$;

$t_{ji}$ matriz dos tempos de viagem do servidor $j$ ao átomo $i$;

$t_{ji}^{[1]}$ matriz dos tempos médios de viagem servidor – átomo para chamadas tipo 1;

$t_{ji}^{[2]}$ matriz dos tempos médios viagem servidor – átomo para chamadas tipo 2;

$t_{ji}^{[m>1]}$ matriz dos tempos médios viagem serv- átomo para chamadas múltiplo despacho($m > 1$);

$\overline{T}_q$ tempo médio de viagem para chamadas em fila;

$\overline{T}$ tempo médio de viagem no sistema;

$\overline{T}_i$ tempo médio de viagem ao átomo $i$;

$\overline{TU}_j$ tempo médio de viagem do servidor $j$;

$\overline{T}^{[1]}$ tempo médio de viagem no sistema para chamadas do tipo 1;

$\overline{T}^{[2]}$ tempo médio de viagem no sistema para chamadas do tipo 2;

$\overline{T}_t^{[2]}$ tempo médio de viagem no sistema para chamadas do tipo 2 (incluindo todos os servidores despachados);

$\overline{T}_F$ tempo médio de viagem para o primeiro servidor a chegar no local de uma chamada tipo 2;

$\overline{T}_S$ tempo médio de viagem para o segundo servidor a chegar no local de uma chamada tipo 2;

$\overline{TU}_j^{[1]}$ tempo médio de viagem para cada servidor $j$, em atendimentos tipo 1;

$\overline{TU}_j^{[2]}$ tempo médio de viagem para cada servidor $j$, em atendimentos tipo 2, nos quais $j$ é o servidor preferencial;

$\overline{T}(c)$ tempo médio de viagem no sistema para uma dada solução $c$;

$\overline{T}_{linit}$ limitantes discretos para $\overline{T}$;

$Ts_j$ tempo de atendimento do servidor $j$ utilizado como dado de entrada do modelo de simulação;

$T_0j$ tempo de atendimento dos dados coletados do sistema e utilizado como dado de entrada no modelo Hipercubo;

$\tau_{ri}$ matriz dos tempos de viagem entre os átomos $r$ e $i$ do sistema;

U.T.I veículo com equipamentos médico emergenciais especializados

$x_i^0$ tamanho inicial do átomo $i$;

$x_i$ tamanho final do átomo $i$;

$y_j$ proporção da distância entre dois servidores adjacentes $j$ e $j+1$ que corresponde a área primaria do servidor $j$;

## Resumo

O objetivo deste trabalho é desenvolver métodos efetivos para analisar a configuração e operação de sistemas de atendimento emergencial (SAEs) em rodovias. Devido às características estocásticas de tais sistemas, principalmente nos processos de chegada e atendimento dos chamados de emergência, aplicamos o modelo Hipercubo para analisar as medidas de desempenho do sistema. Este modelo, conhecido na literatura de localização de sistemas de emergência, é baseado em teoria de filas espacialmente distribuídas. Os SAEs em rodovia operam com uma política de despacho particular, a qual admite que apenas algumas ambulâncias do sistema possam viajar a determinadas regiões (*backup* parcial) e utiliza múltiplo despacho de ambulâncias para atender a certas chamadas. Neste trabalho estendemos o modelo Hipercubo para analisar tais situações. Como o modelo Hipercubo é descritivo, combinamos estas extensões do modelo Hipercubo com um algoritmo genético para obter uma abordagem prescritiva capaz de otimizar a configuração e operação de SAEs em rodovias. Tal abordagem pode ser útil para apoiar decisões no plano estratégico, por exemplo, a localização das bases das ambulâncias ao longo da rodovia e o dimensionamento das regiões de cobertura de cada base. Assim como apoiar decisões no plano operacional, por exemplo, a escolha da política de despacho das ambulâncias para atender chamados de urgência e a determinação das áreas de cobertura de cada servidor (quando a configuração do sistema puder ser alterada de acordo com as condições operacionais de uma semana ou de um dia). Para analisar o desempenho desta abordagem, realizamos estudos de casos com dados reais do sistema *Anjos do Asfalto* (rodovia Presidente Dutra) e da concessionária *Centrovias* (trechos das rodovias Washington Luis, Eng. Paulo Nilo Romano e Comandante João Ribeiro de Barros), no interior de São Paulo. Os resultados mostram que a abordagem é efetiva para apoiar decisões relacionadas ao planejamento e operação destes sistemas.



**Palavras-chave:** modelo Hipercubo, sistemas medico emergencial, despacho de ambulâncias, rodovias

## Abstract:

*The purpose of this study is to develop effective methods to analyze the configuration and operation of the emergency medical systems (EMS) on highways. Due to the stochastic nature of these systems, especially in the arrival and assistance processes of the emergency calls, we apply the Hypercube Queuing Model to evaluate the performance measures of the system. This is a well-known model in the location literature, which is based on spatially distributed queuing theory. The EMS on highways operate within a particular dispatching policy which considers that only some ambulances in the system can travel to certain regions (partial backup) and multiple dispatch of ambulances to respond to certain calls. In this study we extend the Hypercube model to deal with these situations. Since the Hypercube model is a descriptive model, we also develop a Hypercube embedded genetic algorithm to create a prescriptive approach to optimize the configuration and operation of EMS on highways. This approach can support decisions at the strategic level, for example, the location of ambulances along the highway and the primary response area to each ambulance, as well as, decisions on the operational level, for example, the optimal dispatch policy of ambulances to respond to the emergency calls and the coverage area to each ambulance (if the system configuration can be modified according to the operational conditions of the week or the day). In order to evaluate the performance of the proposed approach, we conducted experiments using the data of two real-systems: the EMS Anjos do Asfalto (Presidente Dutra highway) and EMS Centrovias (portions of the highways Washington Luis, Eng. Paulo Nilo Romano e Comandante João Ribeiro de Barros) in São Paulo State. The results show that the approach is effective to support planning and operation decisions in such systems.*

***Keywords:*** *Hypercube model, emergency medical systems, ambulance deployment, highways*.

# 1. Introdução

Sistemas de serviço congestionados deterioram a qualidade de seus serviços em parte devido aos atrasos nos tempos de resposta aos usuários. Embora estes atrasos possam ser tolerados, até certo ponto, em grande parte dos serviços urbanos (como, p.e., numa fila bancária ou na coleta de lixo), existem serviços que não podem admitir maiores demoras, sob o risco de ocorrerem grandes perdas humanas e/ou materiais: são os Serviços de Atendimento Emergencial (SAEs), prestados, por exemplo, pela polícia, bombeiros e resgate médico (ambulâncias).

Dado que, devido a restrições de orçamento, os SAEs não podem ser planejados de forma a trabalhar com um número muito grande de servidores, há claramente um importante *trade-off* a ser considerado entre a qualidade de atendimento e os custos de investimento e operação nestes sistemas. Além disso, como os SAEs são serviços do tipo servidor-para-cliente (*server-to-costumer*), em que os servidores precisam se deslocar até o local da solicitação do usuário, a análise de seu funcionamento em geral, precisa levar em conta fatores probabilísticos na distribuição espacial e temporal dos chamados e servidores.

Nos sistemas de atendimento médico emergencial em rodovias brasileiras, a rapidez no atendimento a um chamado é uma das principais medidas de desempenho, dado que o atraso no tempo de resposta pode resultar em seqüelas e estado de invalidez dos acidentados. Além disso, em muitas situações, pode também significar a diferença entre a vida e morte das vítimas envolvidas. A qualidade dos primeiros socorros prestados pelos SAEs também influenciam demasiadamente no progresso dos tratamentos posteriores no hospital, quando estes são necessários.

Durante os últimos anos, novos SAEs foram instalados e a configuração e operação dos existentes estão sendo revisadas em rodovias brasileiras. Um significante impulso nesta direção é a implementação de um sistema de contrato de concessão adotado por alguns estados brasileiros. Através deste contrato, organizações privadas passam a ser responsáveis pelo planejamento, manutenção e outras operações de atendimento ao usuário nos trechos sob concessão. Desta forma, além de assegurar, através dos contratos, melhor nível de serviço ao

usuário de trechos sob concessão, o governo pode concentrar esforços em melhorar os serviços de atendimento ao usuário de trechos que ainda estão sendo sob sua administração.

Apesar do recente aprimoramento dos sistemas de atendimento ao usuário, o número de acidentes em rodovias brasileiras ainda é preocupante. Além disso, estes sistemas têm sido muito pouco estudados. Por exemplo, na maioria dos SAEs em rodovias, as decisões relacionadas à localização de bases de ambulância ao longo das rodovias não são baseadas nas características do sistema de atendimento médico emergencial, pois são instaladas geralmente ao lado de uma praça de pedágio existente.

Dada a relevância dos SAEs em rodovias, o objeto do presente estudo é propor uma abordagem para otimização da configuração e operação destes sistemas. Algumas das principais contribuições desta abordagem é apoiar decisões relacionadas, por exemplo, em determinar onde localizar os servidores e qual deve ser o tamanho das áreas de cobertura de cada servidor, de forma a minimizar o desbalanceamento das cargas de trabalho entre os servidores e/ou minimizar o tempo de resposta para os usuários. De acordo com BODILY (1978), o balanceamento das cargas de trabalho corresponde a uma das principais medidas de desempenho dos sistemas de atendimento emergencial, embora na maioria dos casos, seja uma medida conflitante com o tempo médio de resposta aos usuários.

Outras medidas de desempenho relevantes para um SAE são a fração de atendimentos realizados fora da área de cobertura de cada servidor e a fração de chamadas atendidas em tempo inferior ao um limite determinado (p.e, fração de chamadas atendidas em tempo inferior à $T$ minutos). Esta última medida tem sido muito utilizada pelos analistas dos SAEs, sendo que, em 1973 o *United States Emergency Medical Services* estabeleceu que 95% dos atendimentos médico emergenciais nos EUA deveriam ser atendidos em tempo inferior à 10 minutos. Em alguns SAEs em rodovias, esta também é uma medida utilizada para avaliar o desempenho do sistema. As concessionárias estão cada vez mais interessadas em elevar o nível de serviço ao usuário com base nestas medidas de desempenho, não só pelas cláusulas do contrato de concessão, mas também pela concorrência entre elas por premiação baseada em seu desempenho.

Ao analisar os SAEs, os fatores probabilísticos relacionados à distribuição temporal e espacial dos servidores e chamadas devem ser considerados, dado que a operação destes sistemas é

caracterizada por incertezas com relação à localização e tempo necessário para atender um determinado chamado. O atraso no tempo de resposta está diretamente relacionado ao conflito entre as variáveis aleatórias da demanda por serviço e as restrições de capacidade do sistema. Assim sendo, estes sistemas podem ser analisados por meio de modelos probabilísticos, que consideram as relações entre a demanda por serviços e o tempo de espera para atendimento aos usuários. Particularmente, pelos motivos citados acima, o modelo Hipercubo proposto em LARSON (1974), baseado em *teoria de filas espacialmente distribuídas*, tem se mostrado como um método preciso e robusto para analisar estes sistemas. Este modelo pode representar as incertezas de um SAE, considerando a identidade dos servidores, assim como a cooperação e/ou iteração entre os mesmos. Por outro lado, o modelo Hipercubo é um modelo essencialmente descritivo e não permite determinar uma configuração ótima (ou próxima da ótima) do sistema de acordo com um critério estabelecido.

Importantes contribuições podem ser encontradas na literatura de estudos sobre modelos de otimização que incorporam os aspectos probabilísticos dos SAEs. No entanto, a maioria destes modelos consideram apenas a aleatoriedade associada à disponibilidade dos servidores e não admitem que há outros aspectos probabilísticos que devem ser considerados na análise. Nos estudos de REVELLE (1989), LOUVEAUX (1993), SWERSEY (1994), OWEN & DASKIN (1998), CHIYOSHI et al. (2000) e BROTCORNE et al. (2003), são revistos os principais modelos de localização para analisar os sistemas de atendimento emergencial, desenvolvidos nas últimas décadas.

Uma direção promissora na análise dos SAE parece ser a integração do modelo Hipercubo em modelos de otimização. São poucas e recentes as publicações com este enfoque, por exemplo: BATTA et al. (1989), SAYDAM & AYTUG (2003), CHIYOSHI et al. (2003), GALVÃO et al. (2003) e GALVÃO et al. (2005). Estes estudos foram bem sucedidos em solucionar os problemas de localização com maior acuracidade e realismo que os modelos anteriores, e são estimulantes para outras pesquisas nesta linha. No presente estudo, estas publicações serviram como base para o desenvolvimento de uma abordagem integrando adaptações do modelo Hipercubo com um algoritmo genético para análise de SAEs em rodovias.

**Contribuições deste estudo:**

Neste estudo é proposta uma abordagem de otimização que integra variações do modelo Hipercubo em um algoritmo genético, para analisar os sistemas de atendimento médico emergencial em rodovias brasileiras. Através desta abordagem é possível prescrever uma configuração suficientemente perto da ótima (senão ótima) do SAE analisado, em termos das principais medidas de desempenho deste sistema. Uma das principais decisões que este método pode apoiar está relacionada em determinar o tamanho dos trechos de rodovia que correspondem à área preferencial (primária) de cobertura de cada servidor. Por exemplo, o método pode apontar qual a melhor configuração em termos do dimensionamento da área preferencial das ambulâncias, de forma a minimizar o desbalanceamento das cargas de trabalho entre as mesmas ou/e minimizar o tempo médio de resposta do sistema, sem que seja necessário investimentos adicionais no sistema. Esta abordagem também permite uma análise do *trade-off* entre estes objetivos, aplicando-se um método simples de otimização bi-objetivo.

Dois estudos de caso são analisados. O primeiro é o SAE *Anjos do Asfalto* na rodovia Presidente Dutra, que foi inicialmente estudado por MENDONÇA & MORABITO (2000, 2001). Para analisar este SAE, por meio do modelo Hipercubo, utilizamos os dados da pesquisa de campo e as informações destes trabalhos. O sistema também foi utilizado como base para a implementação inicial da abordagem de otimização que combina um algoritmo genético com o modelo Hipercubo. Um dos aspectos inovadores desta abordagem é tratar a variação dos tamanhos dos trechos de rodovia, como forma de determinar quais as áreas de cobertura de cada servidor.

O segundo estudo de caso é o SAE de rodovias do interior do Estado de São Paulo, sob administração da concessionária *Centrovias*. Este SAE possui operações muito similares aos SAEs de outras concessionárias de rodovias brasileiras. Porém, diversas características distinguem este sistema do sistema *Anjos do Asfalto*. Uma delas é a política de múltiplo despacho, ou seja, o despacho de duas ambulâncias para atender a determinados tipos de acidente. Além disso, a topologia dos dois sistemas é diferente, pois enquanto que nos *Anjos do Asfalto* todo o trecho é linear (uma única rodovia), no SAE do segundo estudo de caso, os trechos compreendem partes de rodovias radiais e transversais, ou seja, a topologia do trecho total é não-linear. Para analisar este sistema, tomamos por base o trabalho de CHESLT & BARLACH (1981), que propõe uma extensão do modelo Hipercubo para considerar a política

de múltiplo despacho, em um sistema urbano de patrulhamento policial. No presente estudo estendemos e adaptamos esta abordagem para estudar os SAEs em rodovias considerando suas especificidades. Desta forma, novas e interessantes medidas de desempenho podem ser descritas para o sistema, tais como, medidas de freqüências de despacho e tempos de viagem relacionadas às chamadas que requerem duas ambulâncias. Além disso, propomos também uma simples abordagem que integra este modelo em um algoritmo de enumeração exaustiva, de forma a determinar a melhor configuração do sistema em termos das áreas preferências das ambulâncias do sistema. Este método é similar à abordagem que integra o modelo Hipercubo único despacho a um algoritmo genético para análise dos *Anjos do Asfalto*.

Os resultados da aplicação das adaptações do modelo Hipercubo para os dois estudos de caso também foram comparados com os resultados obtidos através de modelos de simulação discreta, para validação das variações dos modelos.

**Estrutura do texto**

Este texto tem sete capítulos, inicia com a presente introdução, e está organizado da seguinte forma:

O segundo capítulo procura discutir as considerações gerais relacionadas aos SAEs, com destaque para os sistemas de atendimento médico emergencial em rodovias brasileiras. O capítulo também apresenta dois estudos de caso de SAEs em rodovias: O sistema *Anjos do Asfalto* inicialmente estudado por MENDONÇA & MORABITO (2000, 2001), e um SAE da concessionária *Centrovias* no interior de São Paulo. Este último tem características típicas dos SAEs de outras concessionárias operando em rodovias no estado de São Paulo e outros estados.

O terceiro capítulo revisa modelos descritivos de localização probabilística desenvolvidos para análise de SAEs, como modelos de Teoria de Filas e simulação. Este capítulo também apresenta, em maior detalhe, o modelo Hipercubo introduzido em LARSON (1974) e algumas extensões que são importantes para tratar as particularidades dos SAEs em rodovias, por exemplo, políticas com único ou múltiplo despacho para chamadas que somente podem ser atendidas por alguns servidores do sistema.

No quarto capítulo propomos modificações no modelo Hipercubo necessárias para sua aplicação nos estudos dos SAEs em rodovias. Analisamos os casos de políticas com único e

múltiplo despacho. Por meio de um exemplo de SAE em rodovias com características similares ao sistema *Anjos do Asfalto*, revisamos as adaptações necessárias para o modelo Hipercubo propostas por MENDONÇA & MORABITO (2000, 2001), considerando que cada região é atendida por determinados servidores do sistema. A seguir propomos adaptações do modelo Hipercubo múltiplo despacho para SAEs em rodovias com características similares ao sistema *Centrovias*, considerando também as particularidades destes sistemas relacionadas à política de despacho, lista de preferência de despacho, diferenciação de servidores e chamadas e localização dos servidores.

O quinto capítulo inicialmente faz uma breve revisão de alguns dos principais modelos prescritivos de localização probabilística que foram desenvolvidos nas últimas décadas para analisar os sistemas de atendimento emergencial. Em seguida, revisa brevemente a metodologia de Algoritmo Genético (GA) e propõe uma abordagem para integração das variações do modelo Hipercubo em um algoritmo genético (GA/Hipercubo), para otimização da configuração e operação de SAEs em rodovias similares aos sistemas *Anjos do Asfalto* e *Centrovias*.

No sexto capítulo, os resultados computacionais deste estudo são apresentados. Estes incluem: (i) a aplicação das adaptações do modelo Hipercubo no sistema *Anjos do Asfalto*; (ii) a aplicação das adaptações do modelo Hipercubo para avaliar o sistema de atendimento emergencial da concessionária *Centrovias*, considerando, entre outros aspectos, a política de múltiplo despacho; (iii) a aplicação da abordagem (GA/Hipercubo) no sistema *Anjos do Asfalto* e em outras instâncias com maior número de servidores geradas aleatoriamente; (iv) a aplicação da abordagem de otimização no sistema *Centrovias*. Nestes últimos, também estudamos o *trade-off* entre as soluções de mínimo desbalanceamento da carga de trabalho dos servidores e mínimo tempo médio de resposta aos usuários. Para isso, aplicamos um método simples de otimização bi-objetivo.

No sétimo capítulo, apresentamos as conclusões deste estudo e apontamos perspectivas para pesquisas futuras.

# 2 - Sistemas de atendimento médico emergencial em rodovias

Este capítulo apresenta uma breve descrição dos sistemas de atendimento emergencial. O foco do capítulo é apresentar as características dos SAEs em rodovias brasileiras. Apresentamos também dois estudos de caso: o sistema *Anjos do Asfalto* na rodovia Presidente Dutra, estudado inicialmente por MENDONCA & MORABITO (2000, 2001) e o sistema de atendimento médico emergencial da concessionária *Centrovias* no interior de São Paulo.

## 2.1 Sistemas de atendimento emergencial (SAEs)

Os SAEs compreendem principalmente os serviços de atendimento emergencial de saúde, os serviços de patrulha policial e os serviços de combate à incêndio. O principal objetivo destes sistemas é garantir o bem estar e segurança da população, oferecendo adequado nível de serviço de forma a evitar perdas humanas e materiais.

A maioria destes sistemas é administrada pelo setor público, como é o caso do corpo de bombeiros, do sistema de atendimento emergencial de saúde e de patrulha policial de uma cidade. Outros podem ser administrados pelo setor privado como os sistemas de segurança pessoal monitorada por uma empresa contratada. Como discutido nas próximas seções, alguns sistemas de atendimento emergencial em rodovias são gerenciados por empresas privadas como parte de um contrato de concessão de trechos das rodovias, inicialmente sob administração do setor público.

Os sistemas de atendimento emergencial são serviços do tipo servidor-para-cliente (*server-to-customer*), nos quais os servidores devem se deslocar até o local do usuário. Uma das principais características destes sistemas é o alto grau de incertezas envolvido, pois há diversos fatores probabilísticos relacionados à distribuição espacial e temporal dos chamados e servidores. Por exemplo, há incertezas acerca do horário, local de uma chamada, e duração de seu atendimento, que aumentam com a complexidade do SAE .

A natureza aleatória da ocorrência de solicitações por atendimento, conflitante com uma capacidade limitada de recursos nos SAEs, pode causar atrasos no tempo de resposta (intervalo de tempo entre a chamada e a chegada do servidor no local da ocorrência). O tempo de espera para receber atendimento em um SAE é uma medida crucial, pois pode trazer

conseqüências graves com perdas humanas e/ou materiais. Desta forma, o tempo de resposta a um chamado de emergência é a principal medida de eficiência destes sistemas.

Como enfatizado por BODILY (1978), além de minimizar o tempo médio de resposta ao usuário, outro importante objetivo do gerenciamento de um SAE é minimizar o desbalanceamento das cargas de trabalho dos servidores do sistema. Como em geral estes dois objetivos são conflitantes, um *trade-off* deve ser considerado na análise e tomada de decisões, com relação aos diferentes interesses envolvidos.

Os três componentes principais de um sistema de atendimento emergencial são:

- Comunicação: em geral, há uma central que deve concentrar as informações do sistema tais como instante de ocorrência e local de uma chamada, estado e local dos servidores, etc. Em geral, o usuário solicita o atendimento e os servidores são enviados ao local para prestar atendimento de acordo com as informações disponíveis e a política de despacho do sistema. Nos dias de hoje, os sistemas de comunicação vêm sendo aprimorados com a introdução, por exemplo, de fibra ótica, monitoramento por câmeras e sistema GPS (*Global Positioning System*).

- Transporte: um SAE deve ser eficiente em chegar no local da ocorrência da forma mais rápida possível transportando pessoal especializado (médico, policial, resgatista, bombeiro) e/ou todo o equipamento necessário (medicamentos, equipamentos de combate ao incêndio). No caso da ambulância e carro resgate, os veículos devem ser adequados para realizar também o transporte de pacientes para o hospital. A localização de bases e chamadas, a política de despacho do sistema, condições de tráfego e o número de servidores disponíveis são os principais fatores que influenciam no tempo médio de viagem. O tempo de viagem corresponde ao intervalo de tempo entre a saída do servidor do local onde se encontra ao receber a ordem de despacho, até a chegada do mesmo no local do atendimento. Note que o tempo de resposta (conforme definido acima) corresponde ao tempo de viagem mais o tempo de preparação (*set-up*) mais um eventual tempo de espera (em fila).

- Atendimento no local: a eficiência do atendimento no local depende da qualificação do pessoal envolvido, dos equipamentos e recursos disponíveis, da organização dos procedimentos e colaboração de terceiros.

A política de despacho de servidores exerce um importante papel nas operações destes sistemas. De acordo com CHAIKEN & LARSON (1972), a política de despacho de um SAE pode ser definida como um conjunto de critérios que estabelecem: (i) o número de servidores de cada tipo enviados para atender uma solicitação por serviço emergencial (este número pode variar de acordo com a hora do dia, dia da semana, ou estação do ano); (ii) a equipe de profissionais enviada; (iii) a localização ou patrulha de cada servidor; (iv) a lista de preferência de despacho para cada tipo de chamada; (v) re-despacho ou re-alocação: sob quais circunstâncias as regras de despacho ou localização dos servidores podem ser alteradas.

Diversos estudos têm sido dedicados a analisar alternativas de políticas de despacho. Apenas citando alguns exemplos temos:
- no despacho de ambulâncias, os estudos de SAVAS (1969) e TAKEDA et al (2000, 2004), que demonstram como o tempo médio de resposta do sistema pode ser substancialmente reduzido se as ambulâncias estiverem dispersas em locais estratégicos, ao invés de centralizadas em um mesmo hospital. TAYLOR & TEMPLETON (1980), EATON et al (1985) e TAKEDA et al (2000, 2004) também analisam a política de despacho em que as ambulâncias despachadas são diferenciadas de acordo com a prioridade do chamado.
- no despacho de viaturas de polícia, os estudos de CHAIKEN & DORMONT (1978a), CHAIKEN & DORMONT (1978b), CHELST (1978), CHELST (1981), CHELST & BARLACH (1981), GREEN (1984), GREEN & KOLESAR (1984a) e GREEN & KOLESAR (1984b) analisam a política de despacho do ponto de vista do número de viaturas enviadas para atender um chamado.
- no despacho de viaturas de bombeiro, os estudos de RIDER (1976), SWERSEY (1982) e IGNALL et al (1982) tratam os aspectos que envolvem a decisão de quantas viaturas de combate a incêndio devem ser enviadas quando ocorre um alarme.

Como comentado em CHAIKEN & LARSON (1972), certos SAEs podem decidir distinguir as chamadas e estabelecer uma política de despacho que não exige servidores especializados, mas evita com que chamadas de alta prioridade espere em fila por atendimento. Por exemplo, quando uma chamada considerada de menor prioridade chega no sistema, o operador de

despacho pode decidir deixá-la em fila ou enviar menor número de servidores para atendê-la, de forma a garantir servidores disponíveis para um possível atendimento de mais alta prioridade.

Uma importante particularidade dos SAEs é a cooperação entre os servidores, pois outros servidores podem também realizar o atendimento a um chamado em uma área, se o servidor mais próximo ou preferencial desta área estiver ocupado. Estes atendimentos são chamados de atendimentos *backup*.

**Características dos principais sistemas de atendimento emergencial**: **bombeiros, patrulhamento policial e ambulâncias.**

Bombeiros

Nos dias de hoje, o serviço prestado pelo corpo de bombeiro não está apenas relacionado aos eventos de combate à incêndios. Na maioria das cidades, as estações de bombeiro também são equipadas com um carro resgate que pode prestar serviços de primeiros socorros. Muitas vezes, estes serviços ocorrem em cooperação com as ambulâncias do sistema emergencial de saúde da cidade. Outras atividades complementares também são realizadas pelo corpo de bombeiros como, por exemplo, o controle e remoção de animais ferozes (cães raivosos) ou venenosos (cobras e escorpiões).

Os veículos despachados para combate à incêndio possuem equipamentos de combate ao fogo, controle de gases tóxicos e salvamento de vítimas. Os veículos resgates possuem equipamentos de primeiros socorros que permitem também condições de transporte da vítima de forma que a mesma possa ser assistida durante a viagem. Como discutido por RIDER (1976), SWERSEY (1994) e IGNALL et al (1982), uma das principais decisões que devem ser tomadas quando um alarme de incêndio é recebido é de quantas e quais viaturas devem ser despachadas, e como devem ser compostas as equipes de profissionais envolvidos.

Patrulhamento policial

Em geral, as viaturas de polícia realizam a patrulha em determinados setores da cidade. Desta forma, estas viaturas estão, na maior parte do tempo, se movendo em seu setor em patrulha. As chamadas são recebidas pela central e comunicadas às viaturas que estão mais próximas do local da ocorrência. Se as viaturas responsáveis pelo setor estão ocupadas, a chamada deve ser atendida por servidores em patrulha nos setores vizinhos. Se a chamada for de baixa prioridade, a mesma pode ainda esperar em fila, de forma que os servidores em outros setores possam estar disponíveis para atender uma possível chamada de mais alta prioridade. As viaturas podem trabalhar em equipe, por exemplo, para os casos que oferecem maior perigo, uma viatura deve oferecer cobertura à outra.

Como descrito em SWERSEY (1994), os principais estudos relacionados ao despacho de patrulha policial são voltados a solucionar os problemas relacionados ao número de viaturas necessárias, determinação dos setores e atribuição de servidores, avaliação do desempenho destes sistemas e programação das equipes de atendimento.

Ambulâncias

De acordo com SWERSEY (1994) e TAKEDA (2000) , desde o início da década de 70 a ambulância deixa de ser apenas um veículo para remoção rápida de vítimas e transporte ao hospital mais próximo. Muitas cidades adquirem ambulâncias equipadas com suporte avançado de atendimento, composto por profissionais especializados, medicamentos e outros equipamentos que permitem o tratamento até mesmo durante o transporte.

O sistema de atendimento emergencial de saúde é oferecido nas cidades brasileiras pelas ambulâncias de hospitais sob a administração das prefeituras municipais. Em geral, as ambulâncias realizam também atendimentos na zona rural e podem ser auxiliadas pelo veículo resgate do corpo de bombeiros. Até poucos anos atrás, as ambulâncias da cidade e o corpo de bombeiros assistiam também acidentes nas rodovias brasileiras. Atualmente, muitas rodovias possuem atendimento emergencial de saúde prestado por empresas privadas como parte do contrato de concessão. Estes SAEs são objeto de estudo deste trabalho e são detalhados nas próximas seções.

Como descrito em TAKEDA (2000) e TAKEDA et al (2000, 2004), desde o início dos anos 90, importantes esforços têm impulsionado melhorias na organização do atendimento médico emergencial no Brasil. Em algumas cidades brasileiras, o modelo de atendimento médico emergencial assemelha-se ao modelo francês SAMU (*Service d'Aide Médicale Urgente*). Os SAMUs são sistemas que operam de forma integrada com um grupo de hospitais públicos em uma cidade ou uma região que engloba várias cidades. O sistema possui uma central que recebe os chamados e realiza o despacho dos veículos e equipe médica necessária de acordo com a disponibilidade dos recursos e a prioridade do chamado. Há dois tipos de ambulâncias: ambulâncias básicas e ambulâncias com suporte avançado de tratamento médico.

O tempo total de atendimento emergencial realizado pelas ambulâncias é composto pelo tempo de recebimento e despacho das ambulâncias, por um possível tempo de espera em fila, pelo tempo de viagem da ambulância ao local do acidente, pelo tempo de atendimento em cena, pelo tempo de transporte ao hospital e pelo tempo de transferência no hospital e o tempo de volta à base ou garagem (no caso de esta não ser no hospital). O atraso no tempo de resposta está diretamente relacionado à sobrevivência dos pacientes.

Os principais estudos dentro da Pesquisa Operacional, voltados para análise dos sistemas de atendimento emergencial, começaram no final da década de 60 e início da década de 70. Os primeiros modelos são relatados em REVELLE et al (1970) e CHAIKEN & LARSON (1972), e as principais referências nesta área nas últimas décadas são encontradas em KOLESAR & SWERSEY (1986), REVELLE (1989), LOUVEAUX (1993), SWERSEY (1994), OWEN & DASKIN (1998), CHIYOSHI et al. (2000) e BROTCORNE et al. (2003).

## 2.2 Os sistemas de atendimento médico emergencial em rodovias brasileiras

Os sistemas de atendimento emergencial (SAEs) em rodovias têm a função de socorrer as vítimas de acidente nas rodovias e, se necessário, realizar o transporte das mesmas ao hospital da cidade mais próxima. Estes sistemas também podem realizar outros tipos de atendimento médico em sua base localizada em um ponto da rodovia.

A importância das operações dos SAEs é fundamental sendo que, infelizmente, o número de acidentes nas rodovias brasileiras ainda é alarmante, e o tempo de resposta médica é

fundamental para a sobrevivência dos acidentados. Os atrasos no tempo de resposta para as vítimas de acidente podem significar a morte, invalidez ou outras seqüelas. Durante os últimos anos, novos SAEs estão sendo implantados e os existentes estão sendo revisados, o que mostra a importância de se pesquisar abordagens para otimização da configuração e operação destes sistemas.

Os SAEs em rodovias são em geral caracterizados por não admitirem fila de espera, pois quando os servidores candidatos estão ocupados, a chamada deve ser transferida à outro sistema como o corpo de bombeiro ou a ambulância da cidade mais próxima, podendo também ser atendida por um outro SAE vizinho. Estes sistemas em geral também admitem políticas particulares de despacho, sendo um que certo servidor do sistema pode nunca ser despachado para atender um determinado chamado em uma certa região da rodovia, devido às restrições de distância. Alguns destes SAEs também possuem uma política de múltiplo despacho, sendo que, em alguns casos (dependendo do tipo de chamada) torna-se necessário despachar mais de um veículo para atender um acidente.

Em geral, as informações são coordenadas por uma central, localizada em um ponto da rodovia ou em uma cidade estratégica. Por meio desta central, as solicitações por atendimento são recebidas e os veículos necessários são despachados de acordo com a política de despacho do sistema. As áreas primárias de cada servidor são determinadas de acordo com os critérios de distância, acesso (sentido, condições geográficas e número de pistas), tipo de acidente e disponibilidade dos recursos.

Os veículos mais utilizados são ambulâncias (resgates) que transportam a equipe de profissionais, equipamentos de primeiros socorros (medicamentos, oxigênio, ressuscitadores, macas e outros) e possivelmente, equipamentos especializados para remoção de ferragens e combate ao fogo ou gases tóxicos. Em rodovias com tráfico relativamente mais intenso, por exemplo, a Rodovia Presidente Dutra (entre São Paulo e Rio de Janeiro), os SAEs contam também com um helicóptero que realiza o resgate em casos de não ser possível a remoção das vítimas do local em tempo hábil pelo veículo resgate. Outros veículos também são comuns como UTI ou carro médico (veja o sistema *Centrovias*, adiante).

A seguir, apresentamos dois estudos de caso de SAEs em rodovias. O primeiro é o sistema *Anjos do Asfalto*, descrito e analisado inicialmente em MENDONCA & MORABITO (2000, 2001), que na seção 2.3 é brevemente revisto. O segundo é o SAE da concessionária

*Centrovias*, como parte de um contrato de concessão dentro do programa de concessões de rodovias do Estado de São Paulo, cuja pesquisa de campo foi realizada no presente estudo.

### 2.3 O sistema *Anjos do Asfalto*:

Como descrito em MENDONÇA & MORABITO (2000, 2001), o sistema *Anjos do Asfalto* é uma fundação não governamental sem fins lucrativos que fornece atendimento médico emergencial em parte da rodovia Presidente Dutra entre as cidades de São Paulo e Rio de Janeiro. Este SAE tem 6 bases fixas ao longo do trecho da rodovia, sendo que, cada base possui uma ambulância e uma equipe composta por médicos, resgatistas, enfermeiros e motorista, que viajam juntos para o local do chamado.

Há uma central de operações, localizada na cidade do Rio de Janeiro, que é responsável por receber as chamadas, despachar as ambulâncias e monitorar os movimentos da mesma. Ao receber um chamado, a central imediatamente envia a ambulância disponível mais próxima do local do chamado. A política de despacho consiste em despachar a ambulância localizada na base mais próxima e, se esta estiver ocupada, a segunda mais próxima é enviada (chamada de *backup*). Se as duas ambulâncias mais próximas estiverem ocupadas, o chamado é transferido para outro sistema (por exemplo, o hospital da cidade mais próxima) e a chamada é considerada uma perda para o sistema. Note que, nesta política de despacho particular, cada região pode ser atendida por somente dois servidores (o servidor preferencial ou o servidor *backup*).

A figura 2.1 ilustra a distribuição das bases dos servidores no sistema *Anjos do Asfalto.* A distância entre duas bases é dividida em duas regiões (ou átomos), cada uma com uma lista de preferência de despacho. De acordo com esta lista, e exceto para as bases das extremidades (1 e 6), todos os servidores são despachados como preferenciais para duas regiões (à esquerda e direita de sua base) e como *backup* para outras duas regiões (direita e esquerda dos servidores adjacentes à esquerda e direita, respectivamente). Note que, o tamanho das regiões varia entre si, por exemplo, as regiões 3 e 4 são relativamente bem menores que as regiões 7 e 8.

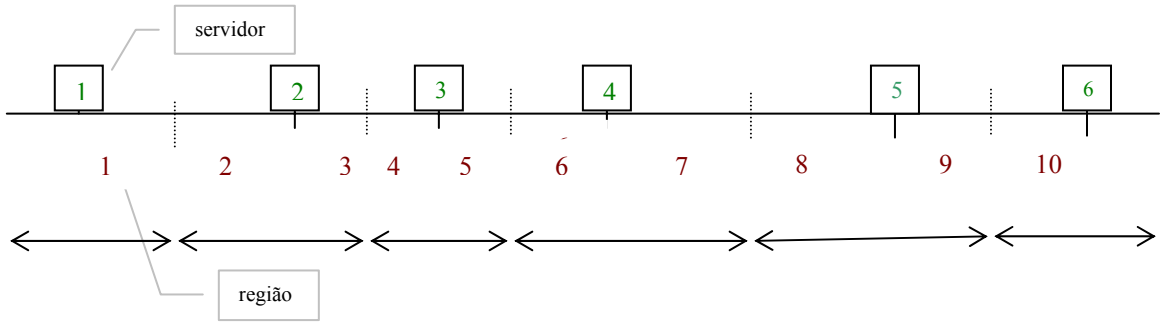


Figure 2.1 – Simples esquema da distribuição de regiões e servidores ao longo da rodovia no sistema Anjos do Asfalto

Por exemplo, para o servidor 5: o lado esquerdo deste servidor (região 8) e o lado direito (região 9) correspondem à sua área primária, ao passo que, o lado direito do servidor 4 (região 7) e o lado esquerdo do servidor 6 (região 10) correspondem à sua área de atendimento *backup*. Os servidores 1 e 6 possuem área primária de uma região (região 1 e região 10, respectivamente), e área *backup* de uma região (região 2 e região 9, respectivamente).

O sistema *Anjos do Asfalto* é analisado com mais detalhes nos próximos capítulos, por meio da adaptação e aplicação do modelo Hipercubo, conforme MENDONCA & MORABITO (2000, 2001). Também analisamos este sistema por meio do algoritmo GA/Hipercubo (Algoritmo Genético combinado ao modelo Hipercubo) para otimizar os tamanhos dos átomos, em função de diferentes critérios de desempenho do sistema.

## 2.4 Os novos SAEs nas rodovias do estado de São Paulo:

### 2.4.1 O programa de Concessões Rodoviárias do Estado de São Paulo

De acordo com o SECTRAN – Secretaria de Estado dos Transportes (2005) e ARTESP – Agência Reguladora de Serviços Públicos Delegados de Transporte do Estado de São Paulo (2005), o Programa de Concessões Rodoviárias do Estado de São Paulo foi autorizado pelo artigo 175 da Constituição Federal e implementado através da lei nº 9.361, de 5 de julho de 1996. Através deste programa foram concedidos 12 lotes de rodovias do Estado, que correspondem a 3.517 km, a 12 empresas privadas.

O contrato de concessão tem duração de 20 anos, durante o qual a empresa se compromete a realizar todas as obras necessárias de planejamento, ampliação e melhoria das rodovias, assim

como prover atendimento ao usuário. Estas empresas são remuneradas através da cobrança de pedágios dos usuários. Segundo dados da ARTESP (2005), até outubro de 2004, o programa de concessões garantiu ao Estado um montante de R$ 1,297 milhões através do pagamento fixo das concessionárias e uma redução de 19,8% no número de mortes por acidentes. As 12 concessionárias do Estado de São Paulo são: *Autoban, Autovias, Centrovias, Ecovias, Intervias, Renovias, Rodovia das Colinas, SPVias, Tebe, Triângulo do Sol, Vianorte* e *Viaoeste.*

O Serviço de Atendimento ao Usuário (SAU), é estabelecido no contrato de concessão e deve oferecer ao usuário: (i) assistência na ocorrência de acidentes; (ii) prestação de socorro médico às vítimas no local e transporte ao hospital se necessário; (iii) apoio na ocorrência de falhas mecânicas e (iv) sistema de informações (0800 e *call box*, sinalização, etc..).

Os trechos administrados pelo DER, que correspondem a 18.000 km, também contam com serviço de atendimento ao usuário. Estes sistemas são chamados de UBAS (Unidades Básicas de Atendimento), e oferecem o mesmo tipo de atendimento oferecido pelos SAUs das empresas privadas.

### 2.4.2 O sistema *Centrovias*

O contrato de concessão da *Centrovias* Sistemas Rodoviários S/A, como parte do programa de concessões do governo do estado de São Paulo, entrou em vigor em junho de 1998. E a partir de setembro de 2002, a *Centrovias* passou a ser administrada pelo grupo OHL (*Obrascon Huarte Laín S/A*), com sede em Madri, Espanha.

Através do contrato de concessão com o DER (Departamento de Estradas de Rodagem), a *Centrovias* passou a administrar trechos da Rodovia SP-310 Washington Luis (Cordeirópolis a São Carlos), Rodovia SP-225 Eng Paulo Nilo Romano (Itirapina a Jaú) e Rodovia SP-225 Comandante João Ribeiro de Barros (Jaú a Bauru).

Na SP-310, o trecho se estende do Km 153 até o Km 227, abrangendo as cidades de Cordeirópolis, Sta Gertrudes, Rio Claro, Itirapina e São Carlos. Na SP –225, o trecho se estende do Km 144 até o Km 235, abrangendo as cidades de Itirapina, Brotas, Dois Córregos, Jaú, Itapuí, Pederneiras e Bauru. De acordo com as especificações do SECTRAN – Secretária

de Estado dos Transportes (2005), a rodovia SP-310 é uma rodovia radial e a SP-225 é uma rodovia transversal.

A *Centrovias* possui SAUs localizados ao longo dos trechos de concessão em cada uma das praças dos seus 5 pedágios. Como mencionado anteriormente, o SAU oferece todo tipo de assistência ao usuário, desde atendimento emergencial de saúde até auxílio mecânico e outros serviços. No caso do serviço médico emergencial, o SAU corresponde também a uma base onde o veículo resgate permanece fixo quando disponível. Cada base também conta com instalações para que os profissionais possam permanecer a espera de um chamado de emergência a qualquer momento. Os cinco SAUs são :

**SAU 1** em Itirapina S.P 310 - Km 217: UTI Médica – Resgate 1

**SAU 2** em Rio Claro S.P 310 - Km 181: Resgate 2

**SAU 3** em Dois Córregos S.P 225 – Km 144: Resgate 3

**SAU 4** em Jaú S.P. 225 – Km 199: Resgate 4

**SAU 5** em Bauru S.P. 225 – Km 225: Resgate 5.

O mapa do trecho total da rodovia administrado pela *Centrovias*, disponível pela ARTESP (2005) é apresentado na figura 2.2 abaixo:

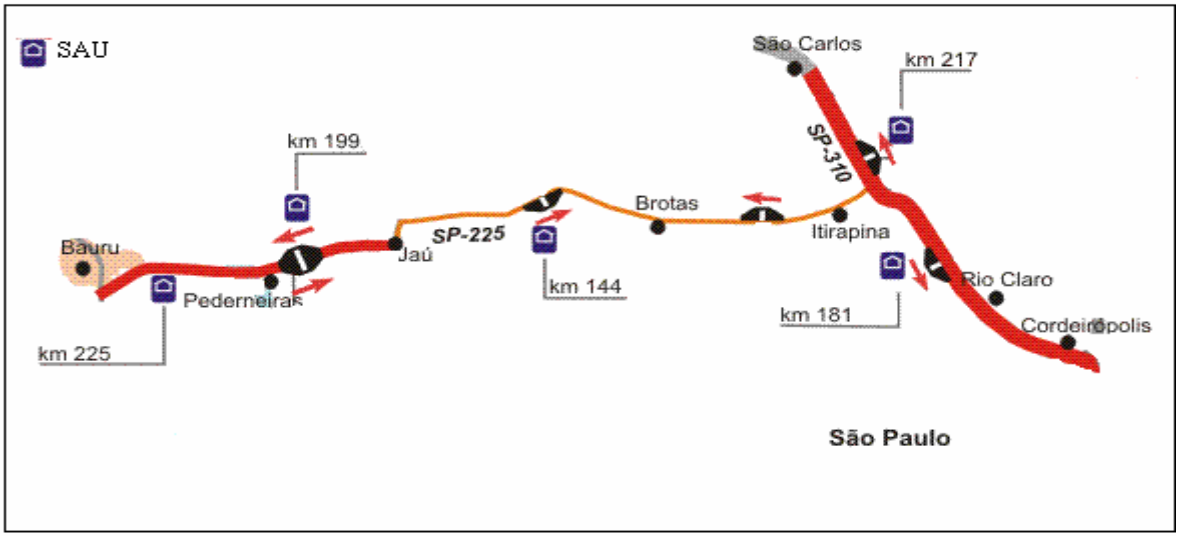


Figura 2.2 : Mapa Centrovias (Fonte: ARTESP –- modificado)

### 2.4.3 O serviço de atendimento médico emergencial dos SAUs:

O sistema de atendimento médico emergencial da *Centrovias* pode atender a todo tipo de ocorrência que requeira assistência médica ou resgate dos usuários que transitam na rodovia, desde um simples mal estar até acidentes muito graves, que envolvem perdas humanas.

Este sistema possui alguns hospitais de referência aos quais são encaminhadas possíveis vítimas de acidentes. Estas unidades também podem auxiliar o atendimento enviando ambulâncias para as rodovias. Os hospitais são: Sta Casa de São Carlos, Sta Casa de Rio Claro, Sta Casa de Jaú, Hospital Base de Bauru e Hospital Benef. Portuguesa de Bauru.

Ao todo o pessoal empregado é de: 1 chefe de setor, 6 médicos e 30 resgatistas (auxiliar de enfermagem e técnico de enfermagem).

Veículos utilizados:

Até o ano de 2003, os veículos do sistema de atendimento médico emergencial na Centrovias eram:

1. U.T.I móvel (1 viatura localizada no SAU 1) composta por: 1 médico (disponível 24 horas), 1 técnico de enfermagem (24 horas) e 1 resgatista.
2. Veículos Resgate (4 resgates, localizados nos demais SAUs) composta por: 1 resgatista e 1 técnico de enfermagem

Apesar desta diferenciação, a U.T.I. atende a qualquer tipo de chamado (grave ou não). A principal diferença deste veículo para os demais é a presença de um médico que monitora os procedimentos realizados pelos demais resgates. Da mesma forma, os resgates podem atender qualquer tipo de acidente, sendo sempre monitorados pelo médico de plantão via rádio. A figura 2.3 ilustra o veículo de resgate.


Figura 2.3 : Resgate (Fonte: CENTROVIAS (2004)).

Central de Chamadas

A central de chamadas, que monitora todo o sistema via rádio e realiza o despacho dos veículos (C.C.O – Central de Controle Operacional), fica localizada no SAU de Itirapina. Os operadores recebem as chamadas de usuários solicitando atendimento nos serviços oferecidos (médico emergencial e mecânico). Uma vez identificados o tipo e local do evento, o operador envia o auxílio necessário.

No caso do atendimento médico, a C.C.O é informada pelo sistema *Centrovias* de informações: via celular ou *call Box* (tel 0800) por um usuário, vítima ou testemunha do acidente. A C.C.O também pode ser notificada da ocorrência de um acidente pelos veículos de inspeção da *Centrovias*, que realizam patrulha contínua ao longo da rodovia. Ao todo são 5 veículos de inspeção : 3 veículos na SP 310 e 2 na SP 225.

Ao receber a chamada, o operador da C.C.O coleta as informações básicas do acidente como local, gravidade e condições correntes. Estas são as informações necessárias para que sejam despachados os recursos de auxílio médico e mecânico necessários. O operador trabalha o tempo todo com um sistema de dados no qual deve atualizar os eventos tanto no caso de acidentes com vítimas ou não. As informações armazenadas são: número do evento, data do evento, horário de acionamento, horário de chegada no local, horário de saída do local, horário de chegada no hospital, horário de saída do hospital e horário de término (retorno da base).

Política de Despacho

Ao receber uma chamada e identificar o local do acidente, o operador da central de despacho deve alocar o resgate do SAU mais próximo. O trecho é dividido em regiões que

correspondem às áreas de atuação primária de cada SAU. Como todos os SAUs estão localizados ao lado de praças de pedágio, a localização dos mesmos não segue nenhum critério específico que esteja relacionado com o atendimento emergencial de saúde. A localização foi estabelecida de acordo com as existentes e novas praças de pedágio. Desta forma, um problema relacionado a este sistema pode ser identificar qual seria a localização ótima dos SAUs do ponto de vista do tempo médio de resposta. Outro problema interessante relacionado à política de despacho deste sistema é determinar qual o tamanho da área de atuação primária de cada SAU. Note que o segundo problema não envolve investimentos adicionais de localização ou relocação de bases. As áreas de atuação primária de cada SAU na configuração atual são:

**SAU 1:** Km 227 – km 195 (SP -310) e Km 91 – km 106 (SP 225);
**SAU 2:** Km 195 – km 153 (SP -310);
**SAU 3:** Km 106 – km 170 (SP - 225);
**SAU 4:** Km 170 – km 215 (SP - 225);
**SAU 5:** Km 215 – km 235 (SP - 225).

(i) Atendimento *backup*

No caso de um acidente ocorrer em um trecho prioritário para um SAU cujo veículo resgate se encontra ocupado atendendo outro chamado, um resgate do segundo SAU mais próximo é despachado. No caso deste também estar ocupado, um sistema auxiliar como o Corpo de Bombeiro é requisitado. O terceiro SAU não é enviado, pois o tempo de viagem resultaria em um tempo de resposta inaceitável para o usuário, devido ao aumento das distâncias. Como o sistema não permite filas, no caso dos dois servidores mais próximos estarem ocupados, a chamada é considerada uma perda para o sistema.

Parte significativa das ocorrências atendidas ocorre no próprio SAU quando usuários solicitam atendimento ao médico e pessoal de plantão, por exemplo, em casos de mal estar, problemas cardíacos, epilepsia, etc

(ii) Múltiplo despacho:

A política de despacho deste SAE também pode enviar mais de um resgate para atender um chamado (os dois resgates mais próximos), dependendo do tipo de chamado. A maior parte de atendimentos *backup* acontece nos trechos que possuem a U.T.I como servidor preferencial ou *backup*. Nestes átomos, podem ocorrer acidentes que requerem equipamentos mais pesados (equipamentos de combate ao incêndio e remoção de ferragens). Como a U.T.I não pode transportar estes equipamentos, a mesma pode receber o apoio do resgate localizado no SAU vizinho. Há outras situações que requerem o envio de dois veículos, por exemplo, em acidentes com mais de duas vítimas, que corresponde a capacidade de cada resgate. No caso da informação sobre o local do acidente não ser precisa (geralmente informação do usuário), também são enviadas duas viaturas.

Quando ocorre um duplo despacho, os dois servidores mais próximos são acionados. Se um destes se encontra ocupado, o servidor livre é despachado e solicita-se o auxílio do hospital ou corpo de bombeiro da cidade mais próxima para transporte das vítimas. O tempo de resposta em um duplo despacho é considerado igual ao do primeiro resgate que chega no local, o qual que deve iniciar imediatamente o atendimento e solicitar reforços.

Np capítulo 4 descrevemos como o modelo Hipercubo pode ser adaptado para analisar SAEs similares ao sistema *Centrovias* 1 (descrito acima). Os resultados da aplicação deste modelo são analisados em detalhes no capítulo 6.

### 2.4.4 O sistema de atendimento médico emergencial da *Centrovias* modificado com a operação do carro médico (*Centrovias* 2).

A partir do ano de 2004, ocorreram algumas mudanças no sistema de atendimento médico emergencial da *Centrovias* com a extinção da U.T.I e aquisição de mais um resgate e um carro médico no SAU 1. Portanto, o sistema passou a operar com 6 servidores, sendo que o carro médico diferencia-se do carro resgate por ser um veículo mais leve que não permite transporte de vítimas, mas que transporta medicamentos, instrumentos básicos de socorro ou operações emergenciais (realização de parto), o médico e o enfermeiro.

Dado que os veículos resgates diferenciam-se do carro médico nos aspectos descritos acima, a política de despacho do sistema também sofreu alterações. Por meio desta nova política, as ambulâncias passam a ser despachadas de acordo com o tipo de chamada de emergência ao longo da rodovia. De acordo com a descrição dos gerentes e operadores do sistema, o despacho dos veículos pode ocorrer nas seguintes formas:

- despacho do carro médico como único despacho: que ocorre para determinados tipos de chamada que não requerem o uso de equipamentos mais especializados e/ou transporte de vítimas;
- despacho de carros resgates como único ou múltiplo despacho, tal como no sistema *Centrovias* 1(anterior). Em geral atendem, por exemplo, acidentes que envolvam transporte de vítimas, quebra de ferragens, combate a incêndio, remoção de vitimas ou animais, entre outros;
- despacho do carro médico e do veículo resgate simultaneamente: diversas chamadas no sistema exigem a presença do médico no local do acidente para realizar ou orientar os procedimentos de emergência necessários, além da operação do veículo resgate, por exemplo, para resgatar as vítimas e realizar o transporte das mesmas;
- despacho de até 3 servidores (carro médico e dois resgates ou três resgates): que ocorre apenas quando o número de vítimas envolvidas no acidente está acima da capacidade de atendimentos de um ou dois resgates.

Como mencionado acima, o carro médico está localizado no SAU1, que corresponde também a base do resgate 1. Desta forma, a maioria dos despachos envolvendo o carro médico e um resgate ocorrem nas regiões preferenciais do SAU1 e SAU2, na rodovia Washington Luis (SP310). Mas podem também ocorrer nas regiões preferenciais do SAU3 na rodovia Eng Paulo Nilo Romano (SP225). Como veremos na análise de dados deste sistema (seção 6.4), nas regiões mais distantes (áreas preferenciais SAU4 e SAU5), o carro médico não é enviado devido as restrições de distância e acesso, mas os procedimentos são monitorados via rádio pelo médico de plantão.

Além disso, a maioria dos despachos de três veículos simultaneamente, também ocorre nas áreas preferenciais dos SAU1, SAU2 e SAU3. Além da maior taxa de acidentes nestes trechos, isso também pode ser explicado se considerarmos que o carro médico não transporta vítimas, e com maior freqüência que em outras regiões, é necessário enviar mais que um resgate para transportar adequadamente as vítimas envolvidas.

Outra diferenciação importante com relação à política de despacho deste novo sistema para o sistema anterior é que quando ocorrem chamadas que requerem o despacho do carro médico e do resgate, e um destes está ocupado, um terceiro veículo pode ser despachado. Por exemplo, quando um acidente ocorre na área preferencial do SAU1 e requer o despacho do carro médico e do resgate 1 e este último está ocupado, o resgate 2 deve ser enviado em dupla com o carro médico, e se o carro médico está também ocupado, o resgate 2 atende a chamada sozinho. Lembre-se que no sistema *Centrovias* 1 (anterior), somente dois servidores podem ser despachos para o local do acidente.

No capitulo 4 descrevemos como o modelo Hipercubo pode ser adaptado para analisar SAEs similares ao sistema *Centrovias* 2, com múltiplo despacho de ambulância, *backup* parcial e servidores diferenciados. No capitulo 6, apresentamos a análise de dados e alguns resultados da aplicação do modelo Hipercubo para descrever este sistema.

# 3 - Modelos Descritivos

Neste capítulo, revisamos modelos descritivos de localização probabilística e/ou configuração do sistema, que foram desenvolvidos para a análise de sistemas de atendimento emergencial. O destaque é para os métodos baseados em simulação e Teoria de Filas. O foco principal deste capítulo é a descrição básica do Modelo Hipercubo de LARSON (1974) e suas extensões que são importantes para análise dos SAEs em rodovias.

Os modelos estritamente descritivos são utilizados na análise de sistemas reais para descrever suas principais medidas de desempenho. No entanto, estes modelos não prescrevem decisões ou configurações, isto é, não apontam uma configuração ideal de operação destes sistemas, como por exemplo, a localização ótima dos servidores ou a melhor política de despacho. Os principais métodos disponíveis utilizados na literatura são baseados em simulação e Teoria de Filas. A principal importância destes métodos é que levam em conta diversas características aleatórias do sistema (incertezas).

No caso de sistemas de emergência onde a probabilidade de ocupação dos servidores é uma medida crítica (por exemplo, ambulância e patrulha policial), as características aleatórias devem ser consideradas. Nestes sistemas, há possibilidade de atraso no fornecimento de serviço devido à natureza aleatória da ocorrência de chamados, conflitante com uma capacidade limitada de atendimento. Além disso, há incertezas acerca do local, horário e duração de atendimento de um chamado, que aumentam com a complexidade do sistema de emergência. No caso de SAEs em rodovias, o atraso no atendimento de um acidente pode estar relacionado à sobrevivência do acidentado.

Diversas referências dos modelos descritivos e probabilísticos aplicados na análise de sistemas de emergência podem ser encontradas em CHAIKEN & LARSON (1972) e SWERSEY (1994). A seguir, apresentamos alguns exemplos.

**Método da raiz quadrada**

O método da raiz quadrada consiste em um método simples de determinar o número de servidores necessários em uma região. Como descrito em LARSON & ODONI (1981) e SWERSEY (1994), o método baseia em estimar o tempo médio de viagem como função do

número de unidades de atendimentos na região, e desta forma encontrar o número de unidades necessárias para que seja atingido o tempo médio de viagem aceitável ou desejado. KOLESAR & BLUM (1973) mostraram que a distância média percorrida por viagem em uma região é, por aproximação, inversamente proporcional à raiz quadrada do número de servidores desta região por unidade de área. A constante de proporcionalidade depende da configuração da região e pode ser determinada através de probabilidade geométrica ou simulação. KOLESAR & BLUM (1973) apontam que uma das principais utilidades do método é descrever os resultados obtidos para o tempo médio de viagem em diferentes políticas de alocação de ambulâncias.

O método da raiz quadrada foi aplicado por IGNALL et al. (1975) de forma eficiente para analisar o sistema do corpo de bombeiros de Nova Iorque, E.U.A, como parte de um grande projeto envolvendo grupo de pesquisadores em Pesquisa Operacional na década de 70, chamado *Rand Fire Project* (SWERSEY, 1994).

**Modelos que utilizam Teoria de Filas**

Com o objetivo de determinar qual número de ambulâncias é necessário em um sistema emergencial de saúde, BELL & ALLEN (1969) utilizaram o modelo de fila $M/G/\infty$. Este modelo admite que as chegadas ocorrem de acordo com um processo de Poisson com taxa $\lambda$, o tempo de atendimento tem uma distribuição geral com média $\frac{1}{\mu}$ e a distribuição do número de ambulâncias ocupadas é Poisson com média $\frac{\lambda}{\mu}$. Os autores consideraram que o número de ambulâncias necessárias deve garantir que a probabilidade de todas as ambulâncias estarem ocupadas seja menor ou igual a uma pequena probabilidade pré-estabelecida. O modelo pode ser aplicado sem que seja necessário especificar a distribuição dos tempos de atendimento, apenas utilizando a taxa média de atendimento para obter as medidas de desempenho desejadas. Como não se trata de um modelo de filas espacialmente distribuídas, os autores consideram que todas as ambulâncias estão localizadas em uma mesma base com a central que recebe os chamados. Como descrito em LARSON & ODONI (1981), o modelo $M/G/\infty$ é uma boa aproximação para análise de sistemas onde a probabilidade de ocupação do sistema é suficientemente pequena, e se comporta probabilisticamente como o modelo $M/M/\infty$ (ambos têm distribuição do número de ambulâncias ocupadas Poisson com média $\frac{\lambda}{\mu}$).

CHAIKEN (1971) utilizou o modelo M/G/$\infty$, de forma similar ao estudo de BELL & ALLEN (1969), para determinar o número necessário de viaturas de bombeiro em um sistema no qual a probabilidade de saturação do sistema não deve atingir um determinado limiar. No caso do sistema de bombeiro, é também necessário considerar que mais de uma viatura pode ser enviada para atender um único chamado, e que o tempo de atendimento de cada viatura deve variar de acordo com a ordem de chegada no local da chamada.

TAYLOR & TEMPLETON (1980) utilizaram um modelo de filas para analisar um sistema emergencial de saúde na cidade de Toronto, Canadá, estabelecendo duas diferentes classes de usuários: (i) aqueles que requerem atendimento emergencial imediato (por exemplo, ataques cardíacos e acidentes de trânsito) e (ii) aqueles que requerem serviços de rotina que podem esperar em fila (por exemplo, transporte entre hospitais). As chamadas por serviço de rotina (com menor prioridade) devem esperar em fila se $N_1$ servidores estão ocupados de forma a disponibilizar $N - N_1$ servidores para atender uma possível chamada por atendimento emergencial imediato. O modelo admite que o processo de chegadas é Poisson com uma taxa única $\lambda$, estabelecendo-se frações distintas para cada uma das duas classes de usuários, e o tempo de atendimento é exponencial com mesma média para as duas classes. As três principais medidas de desempenho descritas pelo modelo são: (i) a probabilidade de uma chamada encontrar um determinado número de ambulâncias disponíveis; (ii) o tempo médio de espera em filas dos chamados com menor prioridade e (iii) a probabilidade de um chamado com menor prioridade esperar em fila por um tempo maior que um valor aceitável. Por meio destas medidas foi possível determinar o número de ambulâncias necessário de forma a garantir o nível de serviço desejado.

CHAIKEN & DORMONT (1978a, 1978b) desenvolveram o modelo de alocação de viaturas em um sistema de patrulhamento policial, chamado (*Patrol Car Allocation Model* - PCAM). O método utiliza o modelo de filas M/M/N (processo de chegada Poisson, processo de atendimento exponencial e N servidores paralelos idênticos no sistema) com classes de prioridade, e calcula os tempos médios de viagem através do método da raiz quadrada. De acordo com os autores, quando o modelo é utilizado para descrever o sistema, entre as principais medidas de desempenho descritas pelo modelo estão o número médio de viaturas disponíveis, a freqüência de patrulha preventiva, o tempo médio de viagem e o tempo médio de espera em fila para cada classe de prioridade. O estudo concluiu que uma significante fração do tempo de operação do sistema é consumida em atividades não relacionadas com o

atendimento de chamadas de emergência, por exemplo, refeições, serviços de reparo e transporte de prisioneiros.

GREEN (1984) propôs um modelo de filas com múltiplo servidores com prioridades, para análise de um sistema de patrulhamento policial considerando a política de múltiplo despacho, o MCD (*Multiple Car Dispatch*). Neste modelo, o número de servidores enviados para atender um evento depende do tipo de evento e da disponibilidade dos servidores. Em GREEN & KOLESAR (1984a), os autores comparam o modelo MCD com o modelo M/M/N com prioridades, cujos parâmetros são ajustados para considerar políticas de múltiplo despacho. O estudo comprova a superioridade das estimativas obtidas pelo MCD. Este também foi utilizado em GREEN & KOLESAR (1984b) para analisar as alternativas de enviar certo número de patrulhas com um policial versus enviar menor número de patrulhas com dois policiais para atender um chamado.

Os modelos descritivos que utilizam a teoria de filas simples (por exemplo, M/M/N, M/G/1, M/G/$\infty$) não são adequados para tratar os servidores individualmente, assim como as variáveis geográficas e temporais ligadas à localização aleatória das chamadas e dos servidores. Além disso, estes modelos não consideram a cooperação e/ou interação entre os servidores, que é característica dos SAEs. Desta forma, os modelos baseados em Teoria de Filas Espacialmente Distribuídas, em particular o modelo Hipercubo de LARSON (1974), e as técnicas de simulação, correspondem aos métodos mais adequados para analisar os sistemas emergenciais, cujos servidores possuem características operacionais distintas e cooperam entre si.

### 3.1 Simulação

Os modelos de simulação permitem analisar os sistemas reais de forma mais detalhada, sem que seja necessário admitir várias simplificações requeridas pelos modelos analíticos. Muitos sistemas são tão complexos que não podem ser avaliados por meio de modelos analíticos, sendo que, nestes casos, a simulação é a ferramenta mais adequada para analisar o sistema e oferecer suporte na tomada de decisões. Além de descrever o comportamento de um sistema existente, uma das principais características da simulação é permitir que configurações alternativas possam ser facilmente testadas considerando, por exemplo, diferentes políticas de operação do sistema, aumento ou redução dos recursos disponíveis, aumento de demanda e

outras modificações. Os resultados das medidas de desempenho de interesse são analisados nos diferentes cenários e comparados de forma a identificar as configurações mais promissoras. Em PEGDEN et al (1995) e BANKS (1998) são apresentados as principais vantagens da simulação e outros aspectos que devem ser considerados para uma adequada análise de um sistema via simulação.

A simulação também tem sido utilizada nos estudos de Pesquisa Operacional como instrumento de validação dos modelos analíticos. Através da simulação pode-se verificar se as simplificações adotadas por estes modelos não comprometem os resultados da análise.

Como discutido em IGNALL et al (1978) e LARSON & ODONI (1981), se comprovado que um modelo analítico pode ser utilizado para análise de um sistema real, o mesmo deve ser utilizado para análises futuras dado que, em geral, a simulação é um método relativamente caro e a interpretação dos resultados estatísticos é em geral mais difícil que nos métodos analíticos. IGNALL et al (1978) enfatizam as vantagens da utilização dos métodos analíticos ao invés da simulação: (i) em geral, os métodos analíticos podem ser incorporados em outros modelos; (ii) os métodos analíticos exigem menor detalhamento e análise dos dados de entrada, a um menor custo que a simulação.

SAVAS (1969) foi um dos pioneiros em considerar as características aleatórias do sistema na análise de localização utilizando simulação. O estudo analisa o sistema de despacho de ambulâncias de um único distrito do Brooklyn, em Nova Iorque. No modelo original de simulação, todas as ambulâncias estão localizadas no hospital do distrito e então alguns cenários são testados localizando algumas ambulâncias em garagens próximas das regiões com maior incidência de ocorrências. Os resultados obtidos mostraram que o tempo de resposta a um chamado pode ser reduzido em até 10%. Neste estudo, o autor também analisa os custos e benefícios de adicionar ambulâncias ao sistema. As melhores alternativas testadas foram aquelas que consideram ao mesmo tempo adicionar e dispersar a ambulâncias em garagens satélites.

Um exemplo de trabalho mais recente utilizando simulação para a análise de sistemas emergenciais é o estudo de ZAKI & CHENG (1997). Os autores apresentam um modelo de simulação para avaliar um sistema de patrulhamento policial em Richmond, Virginia. O modelo também analisa diferentes alternativas de alocação de veículos considerando diversas complexidades do sistema, tais como zonas não-homogêneas de demanda cujo processo de

chegada nem sempre é exponencial, e variação das condições e operação do sistema de acordo com o período de tempo (dia da semana, estação do ano, horas de pico, etc).

Vários estudos utilizando simulação em sistemas de atendimento emergencial estão voltados para validar os modelos analíticos. Em IGNALL et al (1978), os autores utilizam a simulação para validar alguns modelos analíticos simples aplicados aos sistemas de emergência, tais como: o modelo de fila M/M/N na análise de sistema de patrulhamento policial, o modelo da raiz quadrada utilizado por KOLESAR & BLUM (1973) para estimar a distância média de resposta das viaturas de bombeiro, e outros. O estudo mostrou que estes modelos podem ser seguramente utilizados para análise dos sistemas reais analisados, de forma muito mais barata que a simulação. A simulação também foi utilizada para validação de métodos analíticos nos estudos envolvendo SAEs em FITZSIMMONS (1973), KOLESAR & BLUM (1973), IGNALL et al (1982), SWERSEY (1982), GOLDBERG et al (1990) e outros.

### 3.2 O Modelo Hipercubo

O modelo Hipercubo foi desenvolvido por LARSON (1974) e estendido por vários autores para avaliar os sistemas *server-to-customer*. Basicamente, este é um modelo estocástico descritivo que considera as complexidades geográficas e temporais do sistema e é baseado nos resultados de Teoria de Filas Espacialmente Distribuídas e aproximações Markovianas. Desta forma, o modelo pode analisar um sistema com múltiplos servidores considerando que os mesmos podem estar espacialmente distribuídos ao longo da região, e que possuem características operacionais diferenciadas, mas podem cooperar e/ou interagir entre si. O modelo é apresentado com detalhes em LARSON & ODONI (1981).

O nome Hipercubo é derivado do espaço de estados do sistema, que representa os possíveis estados dos servidores. Considerando que há dois estados possíveis para cada servidor: livre (0) ou ocupado (1) em certo instante de tempo, temos então $O(2^N)$ estados para o sistema. Um estado em particular do sistema é representado pela lista de servidores que estão livres e ocupados. Por exemplo, em um sistema com 3 servidores, o estado 101 corresponde ao estado no qual o servidor 2 está livre e os servidores 1 e 3 estão ocupados. Neste caso o espaço de estado é dado pelos vértices de um cubo, como representado na figura 3.1 abaixo. Se o sistema tem mais que três servidores, temos um Hipercubo.

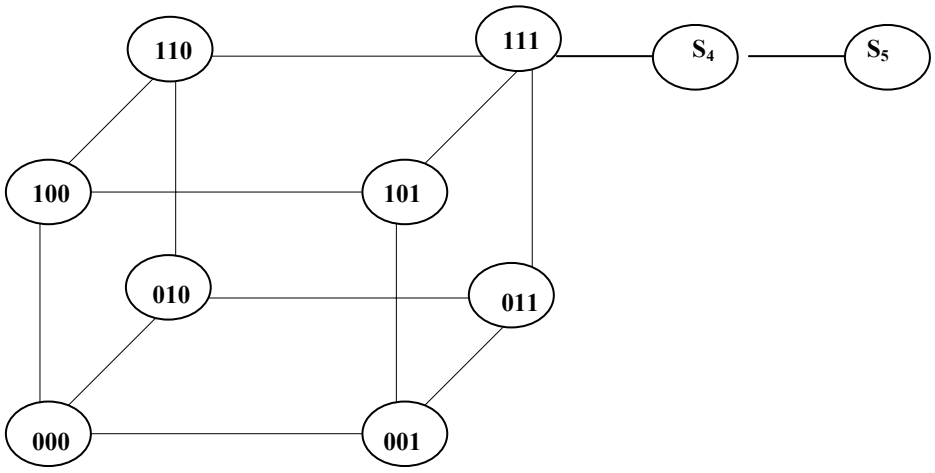


Figura 3.1 – Cubo cujos vértices representam os estados de um sistema com 3 servidores

Observe nesta figura que, a transição de um estado a outro ocorre através das arestas com a mudança de estado de um único servidor, por exemplo, de livre (0) para ocupado (1) e vice e versa. Os estados da cauda $S_4$, $S_5$,.. correspondem aos estados em que todos os servidores estão ocupados (sistema saturado) e há clientes em fila.

A idéia básica do modelo Hipercubo é expandir o espaço de estados de simples modelos de filas com múltiplos servidores (p.e., M/M/N ou M/G/N, onde em ambos, N é o número de servidores), de forma a tratar cada servidor individualmente e incorporar as complexidades das políticas de despacho. O modelo implica na solução de um sistema linear de $O(2^N)$ equações, cujas variáveis envolvidas correspondem às probabilidades de estado do sistema em equilíbrio. Através destas probabilidades, podem ser estimadas importantes medidas de desempenho para análise e gerenciamento do sistema, tais como cargas de trabalho, tempo médio de resposta, frações de despacho de cada servidor a cada região e frações de atendimento fora da área primária. Além de sistemas que consideram fila de espera, o modelo também pode ser aplicado a sistema sem filas, como é o caso de diversos SAEs nas rodovias brasileiras.

A aplicação do modelo Hipercubo implica na divisão da região estudada em átomos, os quais são considerados como fontes solicitadoras de serviço. Os servidores do sistema podem estar fixos ou em movimento ao longo da região, sendo que suas localizações devem ser conhecidas ao menos probabilisticamente. As principais hipóteses do modelo Hipercubo original, descritas por LARSON & ODONI (1981), são:

1. Átomos geográficos: a região em estudo é dividida em $N_A$ átomos geográficos, e cada átomo corresponde a uma fonte independente de chamadas. Desta forma, pode-se considerar as complexidades geográficas e temporais da região.

2. Processo de Chegadas: chamadas de emergência em cada átomo são geradas de acordo com o processo de Poisson, de forma independente dos demais átomos;

3. Servidores: há N servidores espacialmente distribuídos que podem viajar para qualquer átomo;

4. Tempo de viagem: o tempo de viagem entre cada par de átomos é conhecido ou pode ser estimado através de conceitos de probabilidade geométrica;

5. Localização dos servidores: cada servidor, quando livre, pode estar espacialmente distribuído na região de forma estacionária (por exemplo, na base de uma rodovia ou hospital) ou móvel (por exemplo, em patrulha em um dado setor do sistema), e sua localização é conhecida ao menos probabilisticamente.

6. Política de despacho: no atendimento de uma chamada de emergência, exatamente um servidor é despachado para o local da chamada. Se todos os servidores estiverem ocupados, a chamada deve esperar em fila. No caso de sistemas sem filas, a chamada é transferida para outro sistema;

7. Lista de preferência de despacho: o despacho dos servidores é realizado de acordo com uma lista de preferência. Se o primeiro servidor desta lista estiver disponível, o mesmo é despachado, caso contrário, o próximo servidor disponível da lista é despachado (chamado de servidor *backup*). Desta forma a área primária de um dado servidor corresponde ao conjunto de átomos para os quais este servidor é o primeiro a ser chamado. A sua área *backup* corresponde ao conjunto de todos os átomos da região que estão fora de sua área primária, para os quais este servidor é o segundo, terceiro, etc. da lista de despacho. Na maioria dos casos a lista de preferência é determinada com base nas menores distâncias, no entanto, outros critérios podem ser considerados, tais como restrições geográficas e orientação de ruas e estradas, horário e intensidade de tráfego na região, entre outros critérios.

8. Tempo de atendimento: o tempo médio de atendimento para cada servidor é conhecido e inclui o tempo de *set-up* (preparação), o tempo de viagem do servidor ao local da chamada, o tempo em cena e o tempo de retorno do servidor a sua base. Em geral, os servidores possuem tempos de atendimento distintos. O modelo também admite que o desvio padrão dos tempos de atendimento é aproximadamente igual à média (pois o tempo de atendimento é representado por uma distribuição exponencial negativa). Porém, desvios razoáveis desta hipótese não alteram significantemente a precisão do modelo (LARSON & ODONI, 1981). Se o sistema não admite filas, esta suposição é ainda mais desnecessária, pois os modelos M/M/N/N e M/G/N/N têm a mesma distribuição de equilíbrio (CHIYOSHI et al, 2000).

9. Tempo de atendimento dependente do tempo de viagem: variações no tempo de atendimento devido a variações no tempo de viagem são consideradas de segunda ordem, quando comparadas a variações no tempo em cena ou no tempo de *set-up*.

De acordo com SWERSEY (1994) e CHIYOSHI et al (2000), os resultados obtidos na aplicação do modelo dependem de como o sistema estudado se ajusta a estas hipóteses. Algumas das principais limitações do modelo original podem ser apontadas, tais como: o modelo não considera as atividades que, embora não relacionadas com o atendimento de emergência, mantêm os servidores ocupados para o atendimento; o modelo não permite múltiplo-despacho; o modelo não trata prioridades ou diferentes tipos de chamadas. No entanto, diversas extensões do modelo Hipercubo original vem sendo estudadas, por exemplo, em HALPERN (1977), BODILY (1978), CHELST & JARVIS (1979), CHELST & BARLACH (1981), LARSON & MCKNEW (1982), BURWELL et al. (1993) e outros.

O modelo Hipercubo foi originalmente aplicado para análise de sistemas policiais, no qual os servidores podem estar movendo-se em patrulha em uma região dividida em setores. Mas o modelo tem se mostrado eficiente para estudar outros sistemas de emergência. As principais referências de aplicações do modelo Hipercubo aparecem citadas em SWERSEY (1994) e CHIYOSHI et al. (2000), como por exemplo: localização de ambulâncias em Boston (BRANDEAU E LARSON, 1986), o patrulhamento policial em Orlando (SACKS E GRIEF, 1994), e um programa de visitas do serviço social (LARSON & ODONI, 1981). Segundo Larson (2004), recentemente o modelo Hipercubo vem sendo também considerado para aplicação a sistemas de emergência que atuam em caso de ataques terroristas e catástrofes naturais de grande escala (p.e., terremotos, enchentes, maremotos, furações, etc.).

No Brasil, alguns exemplos da aplicação do modelo Hipercubo são: análise de 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), a análise da descentralização de ambulâncias em um sistema derivado do modelo francês SAMU (*Service d'Aide Médicale Urgente*), em Campinas, S.P. (TAKEDA, 2000; TAKEDA et al., 2000, 2004), o estudo do sistema *Anjos do Asfalto* na rodovia Presidente Dutra entre as cidades de São Paulo e Rio de Janeiro (MENDONÇA, 1999; MENDONÇA & MORABITO, 2000, 2001) e a determinação de zonas de atendimento de serviços emergenciais atendidos pelo corpo de bombeiros na cidade de Curitiba, P.R. (COSTA, 2003). Em particular, o trabalho de MENDONÇA & MORABITO (2000, 2001) é revisto no presente estudo.

O modelo Hipercubo permite calcular medidas de desempenho importantes de um sistema a ser analisado. Por exemplo:

1. Medidas para todo o sistema: tempo médio de viagem e/ou resposta no sistema, probabilidade de ocupação do sistema, probabilidade de perda (para sistema sem filas), fração de chamadas em fila (para sistema com filas), tempo médio de espera em fila e número de usuários em fila;
2. Medidas para cada átomo do sistema: tempo médio de viagem e/ou resposta para cada átomo, fração de chamadas do átomo que são perdidas (p.e., no caso de sistemas que não admitem fila de espera), fração de chamadas do átomo que são atendidas por cada servidor, fração de chamadas que são atendidas por servidores *backup*, tempo médio de espera em fila, etc.
3. Medidas para cada servidor do sistema: tempo médio de viagem de cada servidor, fração de despachos para cada átomo, fração de despachos como servidor primário, fração de despacho como servidor *backup*, carga de trabalho, etc.

A seguir, utilizamos um exemplo com três servidores e fila de capacidade infinita para descrever a aplicação do modelo Hipercubo.

**Exemplo ilustrativo 1:**

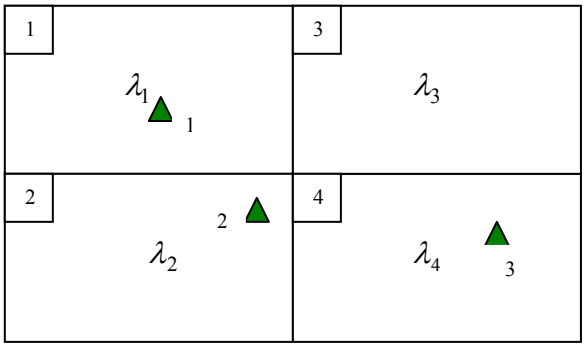


Figura 3.2 : Sistema do Exemplo 1 (sistema com $N_A$ = 4 átomos e N = 3 servidores)

Por meio de um exemplo simples do sistema da figura 3.2 acima, pretendemos apresentar como o modelo Hipercubo original, proposto por LARSON (1974) para a análise de sistemas com fila de capacidade infinita, pode ser aplicado. Neste exemplo, temos:

- Região particionada em $N_A = 4$ átomos (hipótese 1 do modelo Hipercubo);
- O sistema tem N = 3 servidores (representados pelos triângulos na figura 3.2) que quando disponíveis, estão fixos em um dos átomos do sistema (hipótese 5). Neste caso, o servidor 1 está localizado no átomo 1, o servidor 2 no átomo 2 e servidor 3 no átomo 4;
- Em cada átomo *i*, chamadas chegam de acordo com o processo de Poisson com taxa $\lambda_i$, $i = 1,...,4$ (hipótese 2);
- O tempo de serviço é exponencial para cada servidor *j* com taxa de serviço $\mu_j$, $j = 1,...,3$ (hipótese 8);
- A lista de preferência é dada na tabela 3.1 abaixo, considerando que cada servidor pode viajar a qualquer átomo e para cada chamada somente um servidor é despachado (hipóteses 3 e 6). Cada átomo do sistema possui um servidor primário e os demais são considerados servidores *backup* (hipótese 7);
- O sistema admite fila infinita, assim se uma chamada ocorre quando todos os servidores estão ocupados, a mesma entra em uma fila de espera com disciplina FCFS (*first come first server*);
- Os vértices da figura 3.1 representam os possíveis estados deste sistema com até 3 usuários, e que a cauda desta figura representa os estados nos quais há mais de três usuários no sistema, ou seja, quando há usuários em fila;

Tabela 3.1 – Lista de preferência de despacho

| Átomo | Primeiro | Segundo | Terceiro |
|---|---|---|---|
| 1 | 1 | 2 | 3 |
| 2 | 2 | 1 | 3 |
| 3 | 2 | 3 | 1 |
| 4 | 3 | 2 | 1 |

### 3.2.1 Cálculo das probabilidades de estado do sistema:

**Estados do sistema:**

Como mencionado anteriormente, no modelo Hipercubo, os estados do sistema são determinados de acordo com o estado de cada servidor: livre (0) ou ocupado (1). Aqui utilizamos uma representação diferente da utilizada por LARSON (1974). Desta forma, o primeiro digito representa o primeiro servidor, e assim por diante. Por exemplo, para o sistema com três servidores acima temos $2^3$ possíveis estados. Se o estado de um dado servidor $j$ é representado por $b_j$, então um determinado estado do sistema $B$ pode ser representado por : $B = \{b_1, b_2, ....b_N\}$.

No exemplo acima temos os seguintes estados, que correspondem aos vértices da figura 3.1: {000}, {001}, {010}, {011}, {100}, {101}, {110}, {111} e os estados $\{S_4\}$,....{ $S_n$} que correspondem a cauda desta figura. O estado $S_n$ significa que há $n$ usuários no sistema, sendo que ($n$ -3) estão em espera na fila. Analisando o estado $B$ = {011}, temos que neste estado o servidor 1 está livre e os servidores 2 e 3 estão ocupados. De acordo com a lista de preferência de despacho (tabela 3.1), se uma chamada ocorre no átomo 1, a mesma é atendida pelo servidor 1 como área primária, porém se a chamada ocorre nos átomos 2 ou 3, o servidor 1 atende como terceira e segunda preferência, respectivamente.

**Transição de estados:**

Na figura 3.1, observamos que, passar de um estado a outro do sistema corresponde transitar de um dado vértice para os vértices adjacentes, e isto ocorre com o término de serviço (um servidor $j$ passa de ocupado $b_j$ =1 à livre $b_j$ = 0) ou com a chegada de uma chamada (como um único servidor é despachado, um servidor $j$ passa de livre, $b_j$ =0, à ocupado, $b_j$ = 1). Sejam

$\lambda = \sum_{i=1}^{N_A=4} \lambda_i$ e $\mu = \sum_{j=1}^{N=3} \mu_j$ . Discutimos a seguir, como determinar as equações de equilíbrio para diferentes estados do sistema.

A equação de equilíbrio para cada estado $B$ do sistema pode ser construída considerando as transições deste estado para os seus estados (vértices) adjacentes – fluxo para fora do estado e as transições daqueles estados para o estado $B$ – fluxo para dentro do estado.

(i) Quando todos os servidores do sistema estão disponíveis:

No exemplo, o estado $B$ = {000}, corresponde ao estado no qual todos os servidores estão livres. Veja na figura 3.3, que os vértices adjacentes deste estado são: {100}, {010} e {001}. Se $P_B$ corresponde a probabilidade do sistema estar no estado $B$, a equação de equilíbrio do estado $B$ = {000}, pode ser dada por:

$$(\lambda).P_{\{000\}} = \mu_1.P_{\{100\}} + \mu_2.P_{\{010\}} + \mu_3.P_{\{001\}} \qquad (3.1)$$

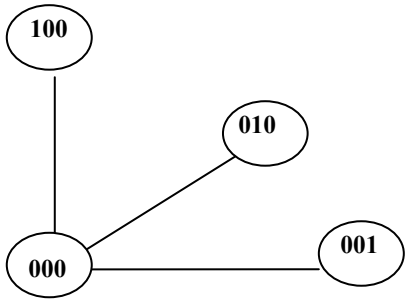


Figura 3.3 – Vértice {000} e seus adjacentes.

Na equação (3.1), o termo do lado esquerdo da igualdade corresponde ao fluxo para fora do estado $B$ = {000}. Note que, a taxa de transição de {000} para {100} é $\lambda_1$, pois como o servidor 1 está livre, uma chamada no átomo 1 deve ser atendida pelo seu servidor primário 1. Assim, como a análise é similar para os estados {010} e {001}, temos que a taxa total de transição para fora do estado {000} é $\lambda$. Por outro lado, como no estado {000}, todos os servidores estão livres, a taxa de transição de algum de seus estados adjacentes para este estado corresponde a taxa de serviço do servidor que está ocupado (probabilidade do servidor ser liberado). Desta forma, a taxa de transição de {100} para {000} é $\mu_1$, e assim por diante

para os demais vértices adjacentes, e o lado direito da equação 3.1 acima corresponde ao fluxo para dentro do estado {000}.

(ii) Quando há 1 servidor ocupado no sistema:

Analisando um outro estado com apenas 1 servidor ocupado, por exemplo, {100}, temos que os vértices adjacentes deste estado são representados na figura 3.4 e a equação de equilíbrio para este estado é dada por:

$$(\lambda + \mu_1).P_{\{100\}} = \lambda_1.P_{\{000\}} + \mu_3.P_{\{101\}} + \mu_2.P_{\{110\}} \qquad (3.2)$$

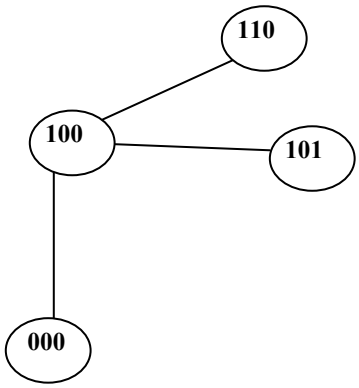


Figura 3.4 - Vértice {100} e seus adjacentes.

Lado esquerdo da equação (3.2) – fluxo para fora do estado {100}: Note na figura 3.4 que as transições para os vértices adjacentes (fluxo para fora ) são:

1. $\{100\} \rightarrow \{101\}$ – ocorre chegada de uma chamada no átomo 4 (taxa $\lambda_4$);

2. $\{100\} \rightarrow \{110\}$ – com a chegada de uma chamada nos átomos 2 e 3 atendidas pelo servidor 2 (como preferencial, veja tabela 3.1), ou com a chegada de uma chamada no átomo 1 (taxa $\lambda_1$), atendida pelo servidor 2 (como *backup,* veja tabela 3.1), dado que o servidor 1 (preferencial) está ocupado. Desta forma, o sistema sai do estado {100}, com chegada de uma chamada em qualquer átomo (com taxa $\lambda$).

3. $\{100\} \rightarrow \{000\}$- que ocorre com o término do serviço do servidor 1 (taxa $\mu_1$).

Lado direito da equação (3.2) – fluxo para dentro do estado {100}:

O sistema passa de {000} para {100} se ocorre uma chamada no átomo 1 (taxa $\lambda_1$), atendida pelo servidor 1. E a transição para {100}, também ocorre com término de serviço dos servidores 2 (taxa $\mu_2$) e 3 (taxa $\mu_3$), nos estados {110} e {101}, respectivamente,

(iii) Quando há 2 servidores ocupados no sistema:

Para um estado com 2 servidores ocupados, por exemplo, {110}, os seus vértices adjacentes são mostrados na figura 3.5 e a equação de equilíbrio é dada por:

$$(\lambda + \mu_1 + \mu_2).P_{\{110\}} = (\lambda_1 + \lambda_2).P_{\{010\}} + (\lambda_1 + \lambda_2 + \lambda_3).P_{\{100\}} + \mu_3.P_{\{111\}} \quad (3.3)$$

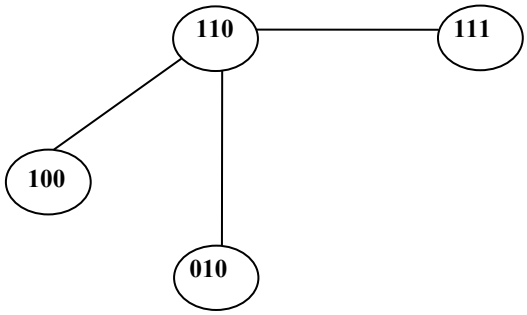


Figura 3.5 - Vértice {110} e seus adjacentes.

Lado esquerdo da equação (3.3) – fluxo para fora do estado {110}: Veja na equação 3.3 e na figura 3.5, que as transições para os vértices adjacentes são:

1. {110} $\rightarrow$ {111} – que ocorre com chegada de uma chamada em qualquer átomo (com taxa $\lambda$) e é atendida pelo servidor 3 (como preferencial ou *backup*, tabela 3.1);
2. {110} $\rightarrow$ {010} – que ocorre com o término de serviço do servidor 1;
3. {110} $\rightarrow$ {100}- que ocorre com o término de serviço do servidor 2;

Lado direito da equação (3.3) – fluxo para dentro do estado {110}: As transições para o vértice {110}, saindo de seus adjacentes (figura 3.5) são:

1. {111} $\rightarrow$ {110} – que ocorre com o término de serviço do servidor 3 (taxa $\mu_3$);
2. {010} $\rightarrow$ {110} - que ocorre com a chegada de uma chamada no átomo 1 (taxa $\lambda_1$) ou no átomo 2 (taxa $\lambda_2$ - atendimento *backup* );

3. $\{100\} \rightarrow \{110\}$ – que ocorre com a chegada de uma chamada nos átomos 1 (taxa $\lambda_1$ - atendimento *backup* ), 2 (taxa $\lambda_2$) e 3 (taxa $\lambda_3$).

(iv) Quando todos os 3 servidores do sistema estão ocupados, estado {111}:

Os vértices adjacentes do vértice {111}, estão representados na figura 3.6, e a equação de equilíbrio é dada por:

$$(\lambda + \mu).P_{\{111\}} = \lambda.P_{\{011\}} + \lambda.P_{\{101\}} + \lambda.P_{\{110\}} + \mu.P_{S_4} \qquad (3.4)$$

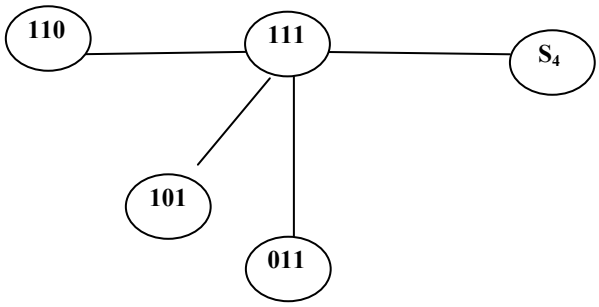


Figura 3.6 - Vértice {111} e seus adjacentes.

Lado esquerdo da equação (3.4) – fluxo para fora do estado {111}: Veja na equação 3.4 e na figura 3.6, que as transições de {111} para os vértices adjacentes de {111} são:

1. $\{111\} \rightarrow \{S_4\}$ – que ocorre com a chegada de uma chamada em qualquer átomo. Lembre que, em $\{S_4\}$ há 4 usuários no sistema (1 em fila e três sendo servidos)
2. $\{111\} \rightarrow \{011\}$ – que ocorre com o término de serviço do servidor 1 (taxa $\mu_1$);
3. $\{111\} \rightarrow \{101\}$ – que ocorre com o término de serviço do servidor 2 (taxa $\mu_2$);
4. $\{111\} \rightarrow \{110\}$ – que ocorre com o término de serviço do servidor 3 (taxa $\mu_3$).

Lado direito da equação (3.4) – fluxo para dentro do estado {111}: As transições para o vértice {111}, saindo de seus adjacentes (figura 3.6) são:

1. $\{S_4\} \rightarrow \{111\}$ – que ocorre com o término de serviço de qualquer servidor, portanto com taxa $\mu$;

2. $\{011\} \rightarrow \{111\}$ – que ocorre com a chegada de uma chamada em qualquer átomo, portanto com taxa $\lambda$, e de acordo com a tabela 3.1, o atendimento é *backup* para os átomos 2, 3 e 4 ;
3. $\{101\} \rightarrow \{111\}$ – que também ocorre com a chegada de uma chamada em qualquer átomo (taxa $\lambda$), e de acordo com a tabela 3.1, o atendimento é *backup* para os átomos 1 e 4;
4. $\{111\} \rightarrow \{110\}$ – como nos dois casos acima, a taxa é $\lambda$, e de acordo com a tabela 3.1, o atendimento é *backup* para os átomos 1,2 e 3;

De acordo com discussão em CHIYOSHI et al (2000) e TAKEDA (2000), ao prosseguir com as equações de equilíbrio para os estados $S_4, S_5, S_6, \ldots$, o resultado seria um sistema infinito de equações. No entanto, as condições de equilíbrio do sistema devem ser respeitadas de forma que as taxas de transição entre os estados $\{111\}$ e $S_4$ devem ser iguais: $(\lambda_1 + \lambda_2 + \lambda_3).P_{\{111\}} = (\mu_1 + \mu_2 + \mu_3).P_{S_4}$. Caso contrário, por exemplo, se $(\lambda_1 + \lambda_2 + \lambda_3).P_{\{111\}} > (\mu_1 + \mu_2 + \mu_3).P_{S_4}$ , o sistema estaria em estado transiente, e a fila estaria em fase de crescimento. Desta forma, a equação de equilíbrio do estado $\{111\}$ pode ser simplificada para:

$$\mu.P_{\{111\}} = \lambda.P_{\{011\}} + \lambda.P_{\{101\}} + \lambda.P_{\{110\}} \qquad (3.5)$$

O sistema de equações de equilíbrio com todos os oito estados - $\{000\}, \{001\}, \{010\}, \{011\}, \{100\}, \{101\}, \{110\}$ e $\{111\}$ é :

$$\{000\}\ \lambda.P_{\{000\}} = \mu_3.P_{\{001\}} + \mu_2.P_{\{010\}} + \mu_1.P_{\{100\}} \qquad (3.6)$$

$$\{001\}\ (\lambda + \mu_3).P_{\{001\}} = \lambda_4.P_{\{000\}} + \mu_2.P_{\{011\}} + \mu_1.P_{\{101\}}$$

$$\{010\}\ (\lambda + \mu_2).P_{\{010\}} = (\lambda_2 + \lambda_3).P_{\{000\}} + \mu_3.P_{\{011\}} + \mu_1.P_{\{110\}}$$

$$\{011\}\ (\lambda + \mu_2 + \mu_3).P_{\{011\}} = (\lambda_2 + \lambda_3 + \lambda_4).P_{\{001\}} + (\lambda_3 + \lambda_4).P_{\{010\}} + \mu_1.P_{\{111\}}$$

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

$$\{101\}\ (\lambda + \mu_1 + \mu_3).P_{\{101\}} = \lambda_1.P_{\{001\}} + \lambda_4.P_{\{100\}} + \mu_2.P_{\{111\}}$$

$$\{110\}\ (\lambda + \mu_1 + \mu_2).P_{\{110\}} = (\lambda_1 + \lambda_2).P_{\{010\}} + (\lambda_1 + \lambda_2 + \lambda_3).P_{\{100\}} + \mu_3.P_{\{111\}}$$

$$\{111\}\ \mu.P_{\{111\}} = \lambda.P_{\{011\}} + \lambda.P_{\{101\}} + \lambda.P_{\{110\}}$$

Note que em todas equações o sistema sai de seu estado corrente para outro com a chegada de uma chamada em qualquer átomo (com taxa total $\lambda$). Isso porque, estamos admitindo que todos os servidores podem atender qualquer átomo (hipótese 6) e que uma chamada pode esperar em fila, caso todos os servidores estão ocupados.

CHIYOSHI et al (2000) explicam que ao tentar resolver o sistema de equações lineares acima na forma matricial A.x = b, percebemos que neste sistema b = 0. Portanto, trata-se de um sistema possível indeterminado pois, ao atribuirmos um valor a uma das probabilidades, podemos determinar as demais a partir desta condição. Isto ocorre porque as equações apenas impõem condições de equilíbrio para cada possível estado (vértice) do sistema ({000},{001},...., {111}), mas nada específica sobre a forma como a massa total de probabilidade se distribui entre estes estados e os estados da cauda ($S_4, S_5, S_6, ...$).

Uma forma de tornar o sistema determinado é a substituição de uma das equações do sistema por uma equação de normalização, considerando que $\sum_{n=0}^{\infty} P_B = 1$ (a soma das probabilidades de todos os possíveis estados do sistema deve ser igual a 1). E a equação de normalização é dada por :

$$P_{000} + P_{001} + P_{010} + ... + P_{111} + P_{S_4} + P_{S_5} + P_{S_6} + .... = 1 \qquad (3.7)$$

No caso de sistema de fila infinita, a equação de normalização pode ser simplificada.Uma simplificação para a equação de normalização é apresentada em CHIYOSHI et al. (2000):

Dado que $\lambda P_{111} = \mu . P_{S_4}$ e $\rho = \frac{\lambda}{\mu}$, onde $\rho$<1, temos que a equação pode ser simplificada para:

$$P_{S_4} = \frac{\lambda}{\mu} P_{111} = \rho . P_{111} \qquad (3.8)$$

E de forma similar para a transição dos estados $S_5, S_6, .....S_\infty$, temos:

$$P_{S_5} = \rho . P_{S_4} = \rho^2 . P_{111} \qquad (3.9)$$

$$P_{S_6} = \rho . P_{S_5} = \rho^3 . P_{111}$$

...

$$P_{S_{k+1}} = \rho.P_{S_k} = \rho^{K-N}.P_{111}$$

....

Assim, somando as probabilidades dos estados nos quais todos os servidores estão ocupados temos:

$$P_{111} + P_{S_4} + P_{S_5} + P_{S_6} + ... = \rho.P_{111} + \rho^2.P_{111} + ... = P_{111}.\sum_{j=0}^{\infty} \rho^j \qquad (3.10)$$

Note que, como $\rho$<1, $\sum_{j=0}^{\infty} \rho^j = \frac{1}{(1-\rho)}$ (uma série geométrica de razão $\rho$) . Portanto,

$$P_{111} + P_{S_4} + P_{S_5} + P_{S_6} + ... = \frac{1}{(1-\rho)} \qquad (3.11)$$

Substituindo na equação de normalização, obtemos:

$$P_{000} + P_{001} + P_{010} + ... + \frac{P_{111}}{(1-\rho)} = 1 \qquad (3.12)$$

### 3.2.2 Medidas de Desempenho

As expressões apresentadas a seguir consideram que os servidores podem não ser homogêneos, ou seja, cada servidor $j$ do sistema tem uma taxa de atendimento $\mu_j$ diferenciada.

Carga de trabalho:

A carga de trabalho de cada servidor *j* é dada pela soma das probabilidades de o mesmo estar ocupado, e pode ser obtida pela expressão:

$$\rho_j = \sum_{\{B:b_j=1\}} P_B + P_q \qquad (3.13)$$

onde $\rho_j$ é a carga de trabalho do servidor *j*, $\sum_{\{B:b_j=1\}} P_B$ é a soma das probabilidades dos estados (de {000} a {111}), em que o servidor *j* está ocupado ($b_j = 1$) e $P_q$ é a probabilidade de haver fila no sistema.

Para o exemplo 1 acima, temos:

$$\rho_1 = P_{100} + P_{101} + P_{110} + P_{111} + P_q \tag{3.14}$$

$$\rho_2 = P_{010} + P_{011} + P_{110} + P_{111} + P_q$$

$$\rho_3 = P_{001} + P_{011} + P_{101} + P_{111} + P_q$$

$$\text{onde } P_q = P_{S_4} + P_{S_5} + P_{S_6} + .... = 1 - (P_{000} + P_{001} + P_{010} + ... + P_{111}) \tag{3.15}$$

Freqüências de despacho:

(i) Freqüência total de despachos do servidor $j$ no átomo $i$.

Uma importante medida de desempenho obtida pelo modelo Hipercubo é a fração de despachos no sistema que são atendidas pelo servidor $j$ no átomo $i$. Definindo $E_{ji}$ como sendo o conjunto dos estados nos quais o servidor $j$ atende um chamado no átomo $i$ (por exemplo, $E_{12} = \left(\{010\},\{011\}\right)$ e considerando que $P_S = P_{111} + P_q$ é a probabilidade de o sistema estar saturado, temos:

$$f_{ji} = f_{ji}^{[nq]} + f_{ji}^{[q]} = \frac{\lambda_i}{\lambda} \sum_{B \in E_{ji}} P_B + \frac{\lambda_i}{\lambda} P_s \frac{\mu_j}{\mu} \tag{3.16}$$

onde o termo $f_{ji}^{[nq]} = \frac{\lambda_i}{\lambda} \sum_{B \in E_{ji}} P_B$ corresponde à fração de todos os despachos em que o servidor $j$ é enviado ao átomo $i$ que não implicam em tempo de espera da chamada na fila, e o termo $f_{ji}^{[nq]} = \frac{\lambda_i}{\lambda} P_s \frac{\mu_j}{\mu}$ corresponde à fração de todos os despachos em que o servidor $j$ é enviado ao átomo $i$ que implicam em algum tempo de espera da chamada na fila . Note que $\sum_{j}^{N} \sum_{i}^{N_A} f_{ji} = 1$.

(ii) Fração de despachos *backup* :

Uma outra medida de fração interessante do sistema é a fração do total de despachos que foram realizados como *backup*, ou seja, fração de chamadas atendidas por outro servidor que não o servidor primário. Esta medida é calculada pela expressão:

$$f_I = \sum_{j=1 i \notin N_{Ai}}^{N} \sum f_{ij} \tag{3.17}$$

onde $f_I$ corresponde à fração do total de despachos que foram realizados como *backup*, $N_{Aj}$ é o conjunto dos átomos para os quais o servidor $j$ é o servidor primário (p.e., $N_{A1} = 1$). No exemplo com N = 3 acima, temos que

$$f_I = f_{12} + f_{13} + f_{14} + f_{21} + f_{24} + f_{31} + f_{32} + f_{33} \tag{3.18}$$

(iii) Fração de despachos do servidor $j$ para átomos fora de sua área primária

Esta é uma medida refere-se ao servidor, e determina a fração total de atendimentos realizados pelo mesmo como servidor *backup*.

$$f_{Ij} = \frac{\sum_{i \notin N_{Ai}} f_{ji}}{\sum_{i=1}^{N_A} f_{ji}} \tag{3.19}$$

onde $f_{Ij}$ é a fração de atendimentos realizados pelo servidor $j$ como *backup*. Por exemplo, a fração de despacho *backup* para servidor $j = 2$, acima é dada por:

$$f_{I2} = \frac{f_{21} + f_{24}}{f_{21} + f_{22} + f_{23} + f_{24}} \tag{3.20}$$

(iv) Fração das chamadas no átomo $i$ que foram atendidos por servidores *backup*:

Esta medida para cada átomo $i$, determina qual a proporção de chamadas que foram atendidas por outros servidores que não seu servidor primário:

$$f_{Ii}^{[a]} = \frac{\sum_{j \neq servidor\_primario\ de\_i} f_{ji}}{\sum_{j=1}^{N} f_{ji}} \tag{3.21}$$

onde $f_{Ii}^{[a]}$ corresponde à fração das chamadas no átomo $i$ que foi atendida por servidores *backup*. No exemplo, para a átomo 2 temos:

$$f_{I2}^{[a]} = \frac{f_{12} + f_{32}}{f_{12} + f_{22} + f_{32}} \tag{3.22}$$

Tempos de viagem:

Uma das principais medidas de desempenho é o tempo de viagem, que está diretamente ligado ao tempo de resposta do sistema. Para determinar as medidas relacionadas ao tempo médio de viagem é necessário utilizar a matriz dos tempos de viagem entre os átomos $r$ e $i$ do sistema ($\tau_{ri}$). Estes valores podem considerar as restrições geográficas do sistema como orientação de ruas e estradas, condições de tráfego. Caso a matriz não possa utilizar dados coletados do sistema, a mesma pode também ser obtida por meio dos conceitos de probabilidade geométrica (hipótese 4). Por exemplo, a matriz pode ser calculada a partir das distâncias entre os centróides de cada átomo (LARSON & ODONI ,1981).

Para os sistemas em que os servidores não estão fixos, por exemplo, patrulha policial, é necessário também determinar a matriz $q_{jr}$ que corresponde a probabilidade de o servidor $j$ estar no átomo $r$. No caso do servidor $j$ estar fixo em um átomo $r$, $q_{jr}$ =1 e a probabilidade deste servidor estar nos demais átomos é nula. A partir destas informações, a matriz dos tempos de viagem do servidor $j$ ao átomo $i$ é calculada por:

$$t_{ji} = \sum_{k=1}^{N_A} q_{jr}.\tau_{ri} \tag{3.23}$$

(i) Tempo médio de viagem para chamadas em fila:

$$\overline{T}_q = \sum_{i=1}^{N_A} \frac{\lambda_i}{\lambda^2} \sum_{j=1}^{N} \sum_{r=1}^{N_A} \frac{\mu_j.\lambda_r.\tau_{ri}}{\mu} \tag{3.24}$$

onde $\frac{\lambda_i}{\lambda}$ é a probabilidade de uma chamada ser gerada no átomo $i$ e $\frac{\mu_j}{\mu}$ é a probabilidade de que o servidor $j$ seja o primeiro a terminar o serviço. Como o sistema está em estado saturado (cauda da figura 3.1), o primeiro servidor a terminar o serviço atende uma chamada e a lista de preferência de despacho de cada átomo não influi no despacho. Portanto, a probabilidade do servidor $j$ estar atendendo no átomo $r$ e viajar deste átomo para $i$ é $\frac{\lambda_r}{\lambda}$, que corresponde a fração de carga de trabalho do sistema gerada no átomo $r$. Note que, como $\sum_{j=1}^{N} \frac{\mu_j}{\mu} = 1$ a expressão (3.24) pode ser simplificada para:

$$\overline{T}_q = \sum_{i=1}^{N_A} \sum_{r=1}^{N_A} \frac{\lambda_i \lambda_r}{\lambda^2} \tau_{ri} \tag{3.25}$$

(ii) Tempo médio de viagem no sistema:

A medida do tempo médio de viagem para uma dada chamada no sistema pode ser obtida por:

$$\overline{T} = \sum_{j=1}^{N} \sum_{i=1}^{N_A} f_{ji}^{[nq]} . t_{ji} + P_s \overline{T}_q \tag{3.26}$$

(iii) Tempo médio de viagem ao átomo $i$ :

$$\overline{T}_i = \frac{\sum_{j=1}^{N} f_{ji}^{[nq]} . t_{ji}}{\sum_{j=1}^{N} f_{ji}^{[nq]}} . (1 - P_s) + \sum_{k=1}^{N_A} \frac{\lambda_r}{\lambda} \tau_{ri} P_s \tag{3.27}$$

(iv) Tempo médio de viagem de cada servidor $j$ :

Segundo LARSON & ODONI (1981) não há uma expressão exata para calcular o tempo médio de viagem para cada servidor $j$ , porém uma boa aproximação é :

$$\overline{TU}_j = \frac{\sum_{i=1}^{N_A} f_{ji}^{[nq]} . t_{ji} + (\overline{T}_q P_S) . \frac{\mu_j}{\mu}}{\sum_{i=1}^{N_A} f_{ji}^{[nq]} + \frac{\mu_j}{\mu} . P_S} \tag{3.28}$$

Outras medidas de desempenho que podem ser obtidas pelo modelo Hipercubo são apresentadas em LARSON (1974) e LARSON & ODONI (1981).

### 3.2.3 Processo de Calibração dos tempos médios de atendimento

Pode ocorrer que em certos sistemas os tempos de viagem representem uma importante parcela no cálculo dos tempos médio de atendimento. Por exemplo, em sistemas de atendimento médico emergencial nos quais a ambulância sai de sua base na cidade para atender uma chamada em área rural, os tempos de viagem da base ao local da chamada, e deste ao hospital ou base, são parcelas relevantes no cálculo do tempo total de atendimento

( $\mu_j^{-1}$ ) de cada servidor $j$ . Para estes sistemas, é necessário ajustar separadamente os tempos médios de viagem de cada servidor de forma a considerar os fatores geográficos que influenciam a viagem de cada servidor.

LARSON & ODONI (1981) descrevem um processo iterativo para calibrar $\mu_j^{-1}$ a partir dos resultados obtidos para $\overline{TU}_j$ (tempo médio de viagem para o servidor $j$ ). O procedimento consiste em verificar a diferença entre a soma de todos os tempos que compõem $\mu_j^{-1}$ (o que inclui $\overline{TU}_j$ ) após a aplicação do modelo, e os valores de $\mu_j^{-1}$ inicialmente utilizados como dados de entrada do modelo. Se a diferença for significativa, o modelo deve ser rodado novamente utilizando os valores de $\mu_j^{-1}$ computados pelo modelo. O processo se repete até que a diferença entre os valores admitidos como dados de entrada do modelo estejam suficientemente próximos dos valores computados pelo modelo.

CHIYOSHI et al. (2000) destacam que seus estudos experimentais com diversos exemplos mostraram que este procedimento costuma convergir em duas ou três iterações, para uma precisão razoável de $\mu_j^{-1}$, embora uma prova que garanta esta convergência não tenha sido encontrada na literatura.

### 3.2.4 Métodos de solução do modelo Hipercubo

Como mencionado anteriormente, as equações de equilíbrio dos estados do sistema resultam em um sistema linear de $2^N$ equações lineares, onde N é o número de servidores. As variáveis a serem determinadas são as probabilidades de estado do modelo Hipercubo. Através destas probabilidades é possível determinar outras medidas de desempenho do sistema. Mesmo para valores moderados de N, é preciso assegurar a viabilidade computacional do modelo, e para certos valores limites de N, é necessário utilizar métodos aproximados de solução.

Método Exato

Para resolver o sistema linear com $2^N$ equações podem ser utilizados o método de Gauss-Jordan ou um método iterativo como os métodos de Gauss-Jacobi ou Gauss-Seidel.

O método de Gauss-Jordan, que é o método utilizado neste estudo, baseia-se no procedimento de eliminação de Gauss. Este é um método exato para determinar a inversa de matrizes, gerando soluções exatas para sistemas lineares do tipo A.x = b.

Os métodos iterativos exigem que se estabeleça um critério de parada, baseado no erro absoluto máximo tolerável e por isso são aproximados por natureza. Por outro lado, como destacado em CHIYOSHI et al. (2000) e CHIYOSHI et al. (2001), estes métodos possuem certas vantagens, como utilizar exclusivamente os elementos não nulos da matriz dos coeficientes, pois envolvem matrizes de coeficientes esparsas. Nestes estudos, os autores apresentam com mais detalhes o uso destes métodos para resolver as probabilidades de estado do modelo Hipercubo utilizando um exemplo com N = 3 servidores, e também discutem resultados de suas experiências computacionais para a solução de problemas testes.

Método aproximado

Como o modelo Hipercubo exato implica na solução de um sistema linear com $2^N$ equações, mesmo para moderados valores de N, o modelo exige um excessivo esforço computacional. Alternativamente, LARSON (1975) desenvolveu um método aproximado para o modelo Hipercubo, o qual apresenta melhor eficiência computacional.

O método introduz fatores de correção para considerar a independência dos servidores, e supõe mesmo tempo de médio de atendimento para todos os servidores. Ao invés das probabilidades de estado, as variáveis do sistema passam a ser as taxas de ocupação dos N servidores do sistema, conseqüentemente o número de equações passa a ser N equações não-lineares. O método é apresentado em detalhes, por exemplo, em LARSON (1975), LARSON & ODONI (1981), CHIYOSHI et al. (2000) e RIVAS (2001).

Os resultados apresentados em LARSON (1975) mostram que os desvios dos resultados obtidos utilizando o método aproximado para sistemas com servidores homogêneos são tipicamente da ordem de 1 a 2%. Apesar do bom desempenho do método, o mesmo só pode ser aplicado para sistemas homogêneos e sistemas com fila de capacidade infinita. Outro aspecto que contribui para o caráter aproximado do modelo é que a derivação dos fatores de correção baseia-a se em um processo de amostragem aleatória como em um sistema de filas

M/M/N, enquanto que no modelo Hipercubo a escolha de um servidor para atender um chamado depende da política de despacho do sistema.

Como discutido a seguir, JARVIS (1985) propôs um algoritmo de aproximação do modelo Hipercubo, para estimar as probabilidades de ocupação dos servidores, relaxando a hipótese de que os mesmos são homogêneos. Em MORABITO et al (2004), os autores discutem os efeitos de considerar os servidores homogêneos versus não-homogêneos utilizando três exemplos com N = 3 servidores para ilustrar três diferentes tipos de sistema. Além disso, a mesma análise é utilizada para dois sistemas reais: o sistema *Anjos do Asfalto* (MENDONCA & MORABITO 2000, 2001) e o SAMU de Campinas (TAKEDA et al. 2000, 2004).

Destacamos a seguir algumas extensões do modelo Hipercubo que são importantes na análise dos SAEs em rodovias, dada as suas particularidades. Outras importantes extensões deste modelo podem ser encontradas em HALPERN (1977), BODILY (1978), CHELST & JARVIS (1979), BURWELL et al. (1993) e SWERSEY (1994).

### 3.2.5 Algumas Extensões do modelo Hipercubo importantes para aplicação em rodovias:

(i) Modelo Hipercubo sem cauda (não admite fila de espera):

Para sistemas que não admitem filas, como é o caso dos SAEs em rodovias, o modelo Hipercubo deve sofrer algumas alterações, sendo que a probabilidade de haver fila é nula e portanto $P_q = 0$ e $\overline{T}_q = 0$. Quando uma chamada chega no sistema e todos os usuários estão ocupados, a mesma não pode ser atendida e é considerada uma perda para o sistema. Assim, uma medida de desempenho interessante do sistema é a probabilidade de perda.

Para o sistema do exemplo 1 com 3 servidores, a cauda da figura não existe e os únicos estados possíveis do sistema são {000}, {001},{010}, {100},{011},{101}, {110} e {111}. Desta forma a equação de normalização é modificada para :

$$P_{000} + P_{001} + P_{010} + ... + P_{111} = 1 \qquad (3.29)$$

Outras expressões também devem ser modificadas com relação aos termos que consideram espera em fila. Por exemplo, no cálculo de $f_{ji}$, o termo $f_{ji}^{[q]}$ torna-se nulo. Este tipo de sistema é discutido com mais detalhes no próximo capítulo utilizando um SAE em rodovia como exemplo de um sistema real que funciona sem a existência de fila de espera.

No estudo de MENDONCA & MORABITO (2000, 2001), os autores estudaram as modificações necessárias para aplicação do modelo Hipercubo na análise de um SAE em rodovia, o sistema *Anjos do Asfalto*. Neste sistema real, a perda de uma chamada é determinada de acordo com uma particular política de despacho estabelecida para o sistema. Este estudo é discutido com mais detalhes nos próximos capítulos deste texto.

(ii) Modelo Hipercubo com fila de capacidade finita.

Em sistemas, nos quais é permitido somente um dado número de chamadas esperando na fila de espera, o número de estados possíveis do sistema também é finito. Note que o caso anterior é um caso particular deste tipo de sistema, pois no caso anterior a capacidade de usuários em fila é igual zero.

Se no exemplo 1 com N = 3 servidores, consideramos que apenas 3 usuários podem permanecer na fila de espera, temos que os possíveis estados do sistema são: {000}, {001},{010}, {100},{011},{101},{110},{111}, $S_4, S_5$ e $S_6$. Considerando as discussões relativas à equação de normalização (3.12), neste caso torna-se:

$$P_{000} + P_{001} + P_{010} + ... + P_{111} + \rho.P_{111} + \rho^2.P_{111} + \rho^3.P_{111} = 1 \qquad (3.30)$$

Nas expressões das medidas de desempenho $P_q \neq 0$ e $T_q \neq 0$, porém há também a probabilidade de perda associada ao sistema, pois chamadas que encontram a fila no limite de capacidade não são atendidas pelo sistema.

Um exemplo de sistema real com fila de capacidade finita é o sistema SAMU de Campinas, S.P, cujo estudo é apresentado em TAKEDA et al. (2000, 2004). Naquele estudo, onde o sistema estudado possui 10 ambulâncias e a capacidade da fila é de 10 chamadas, estas modificações no modelo Hipercubo são abordadas com mais detalhes.

(iii) Modelo Hipercubo para sistemas com prioridades de chamadas

Em alguns sistemas *server-to-customer* é possível que os usuários do sistema sejam atendidos de acordo com um critério de prioridades. Tal característica é comum, por exemplo, em sistemas de atendimento médico emergencial nos quais há ambulâncias especializadas para um certo tipo de acidente. O sistema SAMU de Campinas, estudado por TAKEDA et al.(2000, 2004), é um exemplo deste tipo de sistema. No SAMU algumas ambulâncias são especializadas para atender chamadas com elevada gravidade (VSAs) e outras ambulâncias básicas (VSBs) são despachadas para atender outros tipos de chamados de emergência médica.

Uma das estratégias utilizadas para tratar os critérios de prioridade do sistema é considerar que cada átomo é uma dupla fonte solicitadora de serviço, ou seja, em cada átomo $i$ temos a taxa de chegada $\lambda_{ai}$ que corresponde às chamadas com mais alta prioridade (chamadas avançadas) e a taxa de chegada $\lambda_{bi}$ que corresponde às chamadas com menor grau de prioridade (chamadas básicas). Para as chamadas avançadas as ambulâncias VSA são os servidores primários e as ambulâncias VSB correspondem aos servidores *backup*. De forma similar, para as chamadas básicas, as ambulâncias VSB são servidores primários e as ambulâncias VSA são os servidores *backup*. LARSON & ODONI (1981) denominam este método como "processo de camadas".

Como descrito no capítulo 2, alguns SAEs em rodovias também operam com servidores diferenciados (p.e., veículos resgates, carro médico, helicópteros, etc.), além de chamadas diferenciadas, sendo que dependendo do tipo de chamada em uma região da rodovia é despachado um determinado tipo de veículo (sistemas similares ao sistema *Centrovias 2*). Similarmente ao sistema estudado por TAKEDA et al.(2000, 2004), é preciso considerar que em cada região há chamadas com diferentes listas de preferência de despacho, e assim as regiões (átomos) devem ser subdivididos em camadas ao aplicar o modelo Hipercubo. No capítulo 4, mostramos como este método é combinado com outras adaptações do modelo Hipercubo para analisar o sistema *Centrovias 2*.

(iv) Modelo Hipercubo para sistemas em que os servidores possuem mais de dois estados possíveis:

LARSON & MCKNEW (1982) propuseram uma extensão do modelo Hipercubo, para estudar os sistemas de patrulhamento policial considerando que, uma parte do tempo que permanecem ocupadas, as viaturas de policia estão atendendo eventos que não foram atribuídos por ordem de despacho da central. Os autores chamam estes eventos de PIAs (*Patrol Initiated Activities*), que são geralmente detectados durante a ronda, tais como: advertências à violação de tráfico, vistoria em veículos e residências, perseguição de veículos, assistência a motoristas, entre outros. Os autores desenvolveram a versão exata e aproximada do modelo Hipercubo, assumindo que há três estados possíveis para cada viatura: (i) livre, (ii) ocupada atendendo uma ordem de despacho (CFS – *Call For Service*) e (iii) ocupada atendendo uma PIA. Desta forma, o sistema passa a ter $3^N$ estados possíveis.

Como observado em MENDONÇA & MORABITO (2001), nos SAEs em rodovias, uma ambulância pode ser despachada para atender um chamado durante a viagem de volta a base, após atender um chamado anterior. Neste caso, pode se considerar que cada ambulância apresenta um terceiro estado, e assim pode estar: (i) ocupada atendendo um chamado; (i) livre na sua base e (iii) livre viajando de volta para a base. Portanto, neste caso, também temos $3^N$ estados possíveis para o sistema.

Além disso, nos SAEs em rodovias podem ocorrer atendimentos na base do servidor que não são decorrentes de acidentes ao longo da rodovia. Como no caso das chamadas PIAs, estes atendimentos não envolvem despacho da central de operações, e tempo de viagem, pois o servidor atende na sua base. Como no modelo de LARSON & MCKNEW (1982), a ambulância pode estar em três estados: (i) livre, (ii) ocupada atendendo uma ordem de despacho e (iii) ocupada atendendo um usuário na base. Desta forma, estudamos a extensão desta abordagem na seção 4.3 do capítulo 4.

(v) Método Aproximado do modelo Hipercubo para análise de sistemas considerando servidores não homogêneos:

Como mencionado anteriormente, a solução do modelo Hipercubo exato pode ser inviável em termos computacionais para valores até mesmo moderados de N (número de servidores), pois envolve a solução de um sistema linear com $2^N$ equações. O método aproximado de LARSON (1975) é uma alternativa para análise destes sistemas, pois requer a solução de apenas N equações não lineares. No entanto, uma das principais simplificações deste método é admitir

que os servidores possuem o mesmo tempo médio de atendimento, ou seja, são servidores homogêneos, o que nem sempre é razoável no caso de rodovias.

JARVIS (1985) desenvolveu um algoritmo de aproximação do modelo Hipercubo, baseado no método aproximado de LARSON (1975), para estimar as probabilidades de ocupação dos servidores em sistemas com múltiplos servidores e sem filas. Este algoritmo permite que os servidores sejam tratados de forma distinta em termos do tempo médio de atendimento, considerando também que cada servidor pode ter um tempo médio de serviço para cada átomo do sistema ou para cada tipo de chamada.

Os resultados obtidos por Jarvis mostraram que o método é eficiente, convergindo com um número relativamente pequeno de iterações, e as variações para o modelo exato são relativamente pequenas (em geral, menores que 3%), considerando distribuição do tempo de atendimento *Erlang* ou exponencial. Jarvis também notou que, para $\rho$ muito pequeno, os erros absolutos ficam em torno de 0.2%, pois nestas condições não há quase interação entre os servidores e o método torna-se praticamente exato. Uma limitação do algoritmo é tratar apenas sistemas que não admitem fila, que em geral é o caso dos SAEs em rodovias.

(vi) Modelo Hipercubo para sistemas com política de despacho particular:

Alguns sistemas são caracterizados por uma política de despacho particular, segundo a qual, um dado átomo só pode ser atendido por determinados servidores. Este é o caso da política de despacho dos SAEs em rodovias, nas quais um átomo pode ser atendido por apenas alguns servidores mais próximos (p.e, os dois mais próximos), devido as limitações de distância. No entanto, esta política de despacho contraria a hipótese 3 do modelo Hipercubo original, a qual assume que cada servidor pode atender qualquer átomo.

A extensão do modelo Hipercubo para considerar uma política de despacho particular para análise de SAEs em rodovias foi desenvolvida por MENDONCA & MORABITO (2000, 2001). Esta abordagem também é adotada no presente estudo e discutida em detalhes no próximo capítulo.

(vii) Modelos Hipercubo para análise de sistemas de emergência com política de múltiplo despacho:

CHELST & BARLACH (1981) estenderam o modelo Hipercubo de LARSON (1974) considerando duplo despacho para sistemas que não admitem fila de espera. Esta abordagem é discutida com mais detalhes nesta seção, sendo que no próximo capítulo apresentamos sua adaptação para análise dos SAEs em rodovias, caracterizados por política de múltiplo despacho particular.

Uma das principais hipóteses do modelo Hipercubo original de LARSON (1974) é admitir que para atender a cada chamado por serviço, apenas um servidor é despachado. No entanto, em alguns sistemas de emergência, podem ocorrer eventos que requerem o despacho de dois ou mais servidores para um mesmo chamado. Como por exemplo:

- Nos sistemas de controle de incêndio (corpo de bombeiros), em geral mais de uma viatura é necessária para controlar o incêndio. Além das proporções do incidente, o peso dos equipamentos necessários influencia no número de servidores requeridos.

- Nos sistemas de patrulha policial, dependendo da seriedade da ocorrência, pode ser necessário que as viaturas de polícia trabalhem de forma complementar, por exemplo, uma oferecendo cobertura à outra. Como mencionado em SWERSEY (1994), em certos eventos destes sistemas, enviar duas viaturas equipadas com um policial pode ser mais eficiente que enviar uma viatura equipada com dois policiais.

- Em alguns sistemas de atendimento médico emergencial, pode ser necessário enviar duas ambulâncias para prestar assistência médica, principalmente quando as mesmas são diferenciadas, por exemplo, uma UTI com medicamentos e equipamentos especializados e uma ambulância básica. Em particular, nos sistemas de atendimento emergencial em rodovias (SAUs), o múltiplo despacho deve ocorrer para certos tipos de acidente de trânsito que envolvem um certo número de vítimas além da capacidade de uma ambulância. Além disso, certos acidentes exigem que dois veículos de resgate, diferentemente equipados, trabalhem de forma complementar. A política de múltiplo despacho dos SAUs é discutida no capítulo 4.

- Como destacado em CHELST & BARLACH (1981) e SWERSEY (1994), o estudo das políticas de múltiplo despacho pode ser interessante nos casos em que diferentes sistemas de emergência trabalham de forma integrada. Por exemplo, quando o trabalho do corpo de bombeiros precisa ser auxiliado por uma viatura policial ou por uma ambulância.

Um dos primeiros estudos utilizando o modelo Hipercubo com política de duplo despacho foi o de CHELST (1975), aplicado no patrulhamento policial em New Haven, E.UA. A alternativa utilizada foi dobrar a taxa de chegada das chamadas que requerem dois servidores. Apesar de os efeitos da política de múltiplo despacho refletir-se nas cargas de trabalho dos servidores, esta alternativa não permite que outras importantes medidas de desempenho relacionadas ao caso particular do múltiplo despacho sejam investigadas.

CHELST & BARLACH (1981) propuseram dois modelos: um modelo exato, baseado no modelo Hipercubo exato de LARSON (1974), e um modelo aproximado, baseado no modelo aproximado de LARSON (1975). Estes modelos permitem a análise de sistemas que não admitem fila e nos quais pode ocorrer o despacho de dois servidores idênticos para atender uma única chamada. Os autores aplicaram o modelo para análise de um sistema de patrulha policial em New Haven, considerando dois tipos de chamadas: chamadas tipo 1, que requerem apenas uma viatura policial, e chamadas tipo 2, que requerem duas viaturas. Interessantes medidas de desempenho para este sistema foram introduzidas, por exemplo:

1. Tempo médio de viagem para chamadas do tipo 1;
2. Tempo médio de viagem para chamadas do tipo 2 (incluindo o tempo de viagem de todos os veículos que se dirigem ao local da chamada);
3. Tempo médio de viagem para o servidor preferencial de chamadas do tipo 2;
4. Tempo médio de viagem para o servidor *backup* de chamadas do tipo 2;
5. Tempo médio de viagem para o primeiro servidor a chegar no local de uma chamada do tipo 2;
6. Tempo médio de viagem para o segundo servidor a chegar no local de uma chamada do tipo 2;
7. A média e a distribuição do intervalo de tempo entre a chegada do primeiro e do segundo servidor a chegar no local de uma chamada do tipo 2;
8. A fração de chamadas do tipo 2 que são atendidas pelo servidor $j$ e $k$ simultaneamente;

9. A proporção de chamadas do tipo 2, para as quais o servidor $j$ é o primeiro a chegar no local (quando dois servidores são despachados ao mesmo tempo);

Note que, as medidas 3 e 5 são distintas porque, em alguns sistemas, principalmente nos sistemas em que os servidores estão se movendo quando disponíveis, o primeiro a chegar no local de uma chamada do tipo 2 pode não ser o servidor preferencial. As estatísticas 3 e 4 descrevem o tempo médio de viagem sob o ponto de vista do operador do sistema e as estatísticas 5 e 6 descrevem o tempo médio de viagem sob o ponto de vista do usuário do sistema que espera pela chegada dos dois servidores. A medida 7 é uma importante informação para os sistemas de patrulha policial, pois durante o intervalo de tempo até a chegada da segunda viatura, a primeira viatura pode ficar exposta ao perigo. Como em alguns SAEs em rodovias, uma viatura pode também depender da chegada de outra para realizar o atendimento, dado que trabalham de forma compartilhada, esta medida também pode ser interessante. A medida 9 verifica com qual freqüência, cada servidor do sistema é o primeiro a chegar no local das chamadas do tipo 2.

Hipóteses do Modelo Hipercubo múltiplo despacho (que se distinguem do modelo Hipercubo básico):

- Processo de chegada: em cada átomo do sistema pode ocorrer dois tipos de chamadas com taxas de chegada independentes, de acordo um processo de Poisson, onde $\lambda_i^{[1]}$ corresponde à taxa de chegada de chamadas tipo 1 no átomo $i$ e $\lambda_i^{[2]}$ corresponde à taxa de chegada de chamadas tipo 2 no átomo $i$.

- Processo de atendimento: as chamadas do tipo 1 são atendidas por um servidor *j* cujo tempo de serviço é exponencialmente distribuído com média $1/\mu_j$ e as chamadas do tipo 2 são atendidas por dois servidores $j$ e $k$, com tempo de serviço exponencial independente $1/\mu_j$ e $1/\mu_k$, respectivamente. Note que, com relação ao tempo de atendimento, uma chamada do tipo 2 é tratada como duas chamadas tipo 1 distintas, sendo atendidas cada uma por um servidor. Como discutido em CHELST & BARLACH (1981), tal consideração mantém o mesmo número de estados do sistema do modelo Hipercubo original ($2^N$), pois se diferenciarmos o tempo de atendimento para chamadas tipo 2, teríamos que considerar um estado adicional

para cada servidor (quando está ocupado atendendo uma chamada tipo 2) e o número de estados possíveis para o sistema passa para $3^N$.

- Política de despacho: os servidores são despachados de acordo com uma lista de preferência para cada átomo e o despacho ocorre da seguinte forma:
(i) no atendimento de chamadas do tipo 1, o primeiro servidor ordenado na lista que está disponível é despachado. Se o mesmo estiver ocupado, a chamada é atendida pelo próximo disponível;
(ii) no atendimento de chamadas do tipo 2, os dois primeiros servidores disponíveis ordenados na lista devem ser despachados. Se apenas o último servidor da lista estiver disponível, o mesmo deve ser despachado.

- Fila de espera: os modelos Hipercubo múltiplo despacho exato e aproximado somente tratam sistemas que não admitem fila. Desta forma, quando uma chamada do tipo 1 chega no sistema e todos os servidores estão ocupados, a mesma não é atendida pelo sistema é e considerada como perda. No caso de uma chamada tipo 2, a mesma só é perdida se todos os servidores estiverem ocupados. Se apenas um servidor estiver disponível, o mesmo deve ser despachado para atender a chamada.

No caso de sistemas com múltiplo despacho em que os servidores possuem funções diferenciadas (servidores diferenciados), pode ser necessário considerar mais de uma lista de despacho, pois um servidor ocupado não pode ser sempre substituído por outro. É o caso, por exemplo, das viaturas diferenciadas do corpo de bombeiros, ou de um sistema médico de emergência que possui ambulâncias avançadas e básicas.

**Exemplo ilustrativo 2:**

Para ilustrar o modelo de múltiplo despacho proposto por CHELST & BARLACH (1981), utilizamos um exemplo baseado no exemplo 1. Pretendemos concentrar as discussões no modelo exato, pois no capítulo 4 apresentamos sua adaptação para a análise de um SAE em rodovia, cuja política de despacho inclui despacho de até dois servidores.

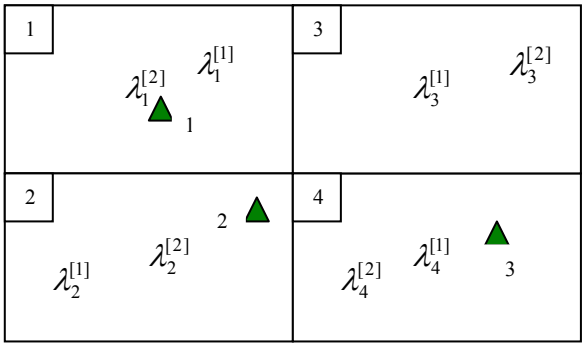


Figura 3.7 : Sistema do Exemplo 2 – múltiplo despacho (com $N_A$ = 4 átomos e N = 3 servidores).

Considere o sistema da figura 3.7, a principal diferença deste sistema com relação ao exemplo 1 é que dois tipos de chamadas, chamadas tipo 1 com taxa de chegada $\lambda_i^{[1]}$ e chamadas tipo 2 com taxa de chegada $\lambda_i^{[2]}$, são geradas em cada átomo $i$. Há $N_A$ = 4 átomos e N = 3 servidores no sistema, sendo que não é permitida fila de espera no sistema, isto é, chamadas que encontram todos os servidores ocupados não são atendidas pelo sistema. A lista de despacho é dada na tabela 3.2. Neste caso, para as chamadas tipo 2, os dois primeiros servidores são enviados. Por exemplo, se uma chamada tipo 2 ocorre no átomo 2, os servidores 2 e 1 são enviados para atendê-la. Se o servidor 2 estiver ocupado, os servidores 1 e 3 são despachados. Se somente o servidor 3 está disponível, o mesmo deve ser despachado sozinho.

Tabela 3.2 – Lista de preferência de despacho

| Átomo | Primeiro | Segundo | Terceiro |
|---|---|---|---|
| 1 | 1 | 2 | 3 |
| 2 | 2 | 1 | 3 |
| 3 | 2 | 3 | 1 |
| 4 | 3 | 2 | 1 |

**Transição de estados:**

Os possíveis estados do sistema são: {000}, {001}, {010}, {011}, {100}, {101}, {110}, {111} que representam os vértices da figura 3.8:

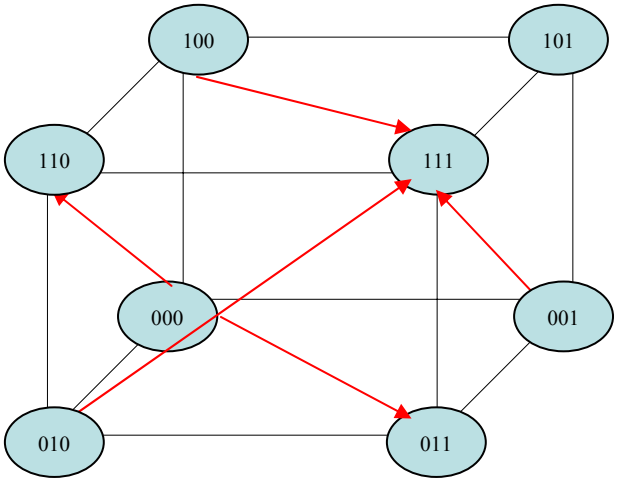


Figura 3.8: Cubo representando os possíveis estados do sistema

Note na figura 3.8 que as setas nas diagonais representam as novas transições de estado que devem ser consideradas, são elas: $\{000\} \rightarrow \{011\}$, $\{000\} \rightarrow \{110\}$, $\{100\} \rightarrow \{111\}$, $\{010\} \rightarrow \{111\}$ e $\{001\} \rightarrow \{111\}$. Note que, a transição diagonal $\{000\} \rightarrow \{101\}$, não ocorre, pois de acordo com a lista de despacho, os servidores 1 e 3 só são enviados juntos se o servidor 2 estiver ocupado, estado $\{010\}$.

**Equações de equilíbrio:**

Nas equações de equilíbrio, temos que $\lambda^{[1]} = \sum_{i=1}^{Na} \lambda_i^{[1]}$, $\lambda^{[2]} = \sum_{i=1}^{Na} \lambda_i^{[2]}$, $\lambda = \lambda^{[1]} + \lambda^{[2]}$ e $\mu = \sum_{j=1}^{N} \mu_j$.

(i) Quando todos os servidores estão disponíveis: $\{000\}$

Note na figura 3.8 que, os vértices adjacentes do vértice $\{000\}$ são $\{100\}$, $\{010\}$ e $\{001\}$. Como discutido anteriormente, para modelo Hipercubo original as transições ao longo das arestas ocorrem nos dois sentidos (fluxo para dentro e para fora de um vértice). No modelo de múltiplo despacho, ocorrem também transições diagonais, pois o sistema também pode ir do estado $\{000\}$ aos estados $\{110\}$ e $\{011\}$. Porém, o sistema não sai destes estados diretamente para o estado $\{000\}$, dado as considerações de processo de atendimento independente entre os servidores. Desta forma, para o estado $\{000\}$, a equação de equilíbrio é:

$$(\lambda).P_{\{000\}} = \mu_1.P_{\{100\}} + \mu_2.P_{\{010\}} + \mu_3.P_{\{001\}} \qquad (3.31)$$

Lado esquerdo da equação (3.31) – fluxo para fora do estado {000}: Analisando as possíveis transições temos:

1. $\{000\} \rightarrow \{100\}$ – ocorre com a chegada de uma chamada do tipo 1 no átomo 1;
2. $\{000\} \rightarrow \{010\}$ – ocorre com a chegada de uma chamada do tipo 1 nos átomos 2 e 3;
3. $\{000\} \rightarrow \{001\}$ – ocorre com a chegada de uma chamada do tipo 1 no átomo 4;
4. $\{000\} \rightarrow \{110\}$ – ocorre com a chegada de uma chamada tipo 2 nos átomos 1 e 2;
5. $\{000\} \rightarrow \{110\}$ – ocorre com a chegada de uma chamada tipo 2 nos átomos 3 e 4;

Desta forma, uma chamada de qualquer tipo e em qualquer átomo transfere o sistema para fora do estado {000}, e a taxa total de transição é $\lambda$.
Lado direito da equação (3.32) – fluxo para dentro do estado {000}:

O sistema passa dos estados {100}, {010} e {001} para {000} com o término de serviço nos servidores 1, 2 e 3, respectivamente.

(ii) Quando há 1 servidor ocupado no sistema:

Analisando por exemplo, o estado {100}, a equação de equilíbrio é:

$$(\lambda + \mu_1).P_{\{100\}} = \lambda_1^{[1]}.P_{\{000\}} + \mu_2.P_{\{110\}} + \mu_3.P_{\{101\}} \qquad (3.32)$$

Lado esquerdo da equação (3.32) – fluxo para fora do estado {100}:

A transição $\{100\} \rightarrow \{000\}$ – ocorre com o término de serviço do servidor 1 (taxa $\mu_1$) e a chegada de uma chamada de qualquer tipo em qualquer átomo transfere o sistema para fora do estado {100}, e a taxa total é $\lambda$.

Lado direito da equação (3.32) – fluxo para dentro do estado {100}: Veja na figura 3.8, que as possíveis transições para dentro deste estado são:

1. $\{000\} \rightarrow \{100\}$ – ocorre a chegada de uma chamada tipo 1 no átomo 1 (taxa $\lambda_1^{[1]}$);
2. $\{110\} \rightarrow \{100\}$ – ocorre com o término de serviço do servidor 2 (taxa $\mu_2$);
3. $\{101\} \rightarrow \{100\}$ – ocorre com o término de serviço do servidor 3 (taxa $\mu_3$).

(iii) Quando há 2 servidores ocupados no sistema:

Note que, de acordo com a figura 3.8, devemos considerar que o sistema pode passar a ter dois servidores ocupados, a mais, simultaneamente. Analisando por exemplo, o estado {110}, a equação de equilíbrio é:

$$(\lambda + \mu_1 + \mu_2).P_{\{110\}} = (\lambda_1^{[2]} + \lambda_2^{[2]}).P_{000} + (\lambda_1^{[1]} + \lambda_2^{[1]}).P_{\{010\}} + (\lambda_1^{[1]} + \lambda_2^{[1]} + \lambda_3^{[1]}).P_{\{100\}} + \mu_3.P_{\{111\}} \quad (3.33)$$

Lado esquerdo da equação (3.33) – fluxo para fora do estado {110}:

1. $\{110\} \rightarrow \{111\}$ – ocorre com a chegada de uma chamada de qualquer tipo em qualquer átomo (taxa total de transição $\lambda$). No caso de chamadas do tipo 2, o atendimento é realizado pelo único servidor disponível;
2. $\{110\} \rightarrow \{010\}$ – ocorre com o término de serviço do servidor 1 (taxa $\mu_1$);
3. $\{110\} \rightarrow \{100\}$ – ocorre com o término de serviço do servidor 2 (taxa $\mu_2$).

Lado direito da equação (3.33) – fluxo para dentro do estado {110}:

1. $\{111\} \rightarrow \{110\}$ – ocorre com o término de serviço do servidor 3 (taxa $\mu_3$);
2. $\{010\} \rightarrow \{110\}$ – ocorre com a chegada de uma chamada tipo 1 no átomo 1 (taxa $\lambda_1^{[1]}$) ou no átomo 2 (taxa $\lambda_2^{[1]}$ - atendimento *backup*);
3. $\{100\} \rightarrow \{110\}$ – ocorre com a chegada de uma chamada do tipo 1 nos átomos 1, 2 ou 3 (taxa total $\lambda_1^{[1]} + \lambda_2^{[1]} + \lambda_3^{[1]}$);
4. $\{000\} \rightarrow \{110\}$ – ocorre com a chegada de uma chamada do tipo 2 nos átomos 1 ou 2 (taxa total $\lambda_1^{[2]} + \lambda_2^{[2]}$ ).

(iv) Quando todos os 3 servidores estão ocupados, estado {111}

Veja na figura 3.8 que, três novas transições ocorrem para dentro deste estado: $\{100\} \rightarrow \{111\}$, $\{010\} \rightarrow \{111\}$ e $\{001\} \rightarrow \{111\}$, com a chegada de uma chamada tipo 2 no sistema. A equação de equilíbrio deste estado, considerando que este é um sistema que não admite fila de espera, é dada por:

$$\mu.P_{\{111\}} = \lambda^{[2]}.(P_{\{100\}} + P_{\{010\}} + P_{\{001\}}) + \lambda.(P_{\{011\}} + P_{\{101\}} + P_{\{110\}}) \quad (3.34)$$

Lado esquerdo da equação (3.34) – fluxo para fora do estado {111}:

Dado que todos os servidores estão ocupados, o sistema sai do estado {111} com o término de serviço de qualquer servidor, e a taxa total é $\mu$.

Lado direito da equação (3.34) – fluxo para dentro do estado {111}:

As transições que ocorrem com chegada de uma chamada do tipo 2, quando há dois servidores disponíveis, para o estado {111} são: $\{100\} \rightarrow \{111\}$, $\{010\} \rightarrow \{111\}$ e $\{001\} \rightarrow \{111\}$, portanto a taxa total de transição é $\lambda^{[2]}$.

O sistema passa dos estados {011}, {101} e {110} para {111} com a chegada de uma chamada tipo 1 ou tipo 2 em qualquer átomo, no caso de chamadas tipo 2, o único servidor disponível é enviado e a chamada não é perdida.

O sistema de equações de equilíbrio para os possíveis estados: {000},{001}, {010},{011},{100}, {101},{110} e {111} é :

$$\{000\}\ (\lambda).P_{\{000\}} = \mu_1.P_{\{100\}} + \mu_2.P_{\{010\}} + \mu_3.P_{\{001\}} \qquad (3.35)$$

$$\{001\}\ (\lambda + \mu_3).P_{\{001\}} = \lambda_4^{[1]}.P_{\{000\}} + \mu_2.P_{\{011\}} + \mu_1.P_{\{101\}}$$

$$\{010\}\ (\lambda + \mu_2).P_{\{010\}} = (\lambda_2^{[1]} + \lambda_3^{[1]}).P_{\{000\}} + \mu_1.P_{\{110\}} + \mu_3.P_{\{011\}}$$

$$\{011\}\ (\lambda + \mu_2 + \mu_3).P_{\{011\}} = (\lambda_3^{[2]} + \lambda_4^{[2]}).P_{000} + (\lambda_2^{[1]} + \lambda_3^{[1]} + \lambda_4^{[1]}).P_{\{001\}} + (\lambda_3^{[1]} + \lambda_4^{[1]}).P_{\{010\}} + \mu_1.P_{\{111\}}$$

$$\{100\}\ (\lambda + \mu_1).P_{\{100\}} = \lambda_1^{[1]}.P_{\{000\}} + \mu_3.P_{\{101\}} + \mu_2.P_{\{110\}}$$

$$\{101\}\ (\lambda + \mu_1 + \mu_3).P_{\{101\}} = \lambda_1^{[1]}.P_{\{001\}} + \lambda_4^{[1]}.P_{\{100\}} + \mu_2.P_{\{111\}}$$

$$\{110\}\ (\lambda + \mu_1 + \mu_2).P_{\{110\}} = (\lambda_1^{[2]} + \lambda_2^{[2]}).P_{000} + (\lambda_1^{[1]} + \lambda_2^{[1]}).P_{\{010\}} + (\lambda_1^{[1]} + \lambda_2^{[1]} + \lambda_3^{[1]}).P_{\{100\}} + \mu_3.P_{\{111\}}$$

$$\{111\}\ \mu.P_{\{111\}} = \lambda^{[2]}.(P_{\{100\}} + P_{\{010\}} + P_{\{001\}}) + (\lambda).(P_{\{011\}} + P_{\{101\}} + P_{\{110\}})$$

Para resolver o sistema acima podemos substituir uma das equações acima pela equação de normalização:

$$P_{000} + P_{001} + P_{010} + ... + P_{111} = 1 \qquad (3.36)$$

**Medidas de Desempenho:**

Os principais medidas de desempenho apresentadas a seguir são relacionadas com a carga de trabalho dos servidores, frações de despacho e tempos de viagem dos atendimentos de chamadas tipo 2.

Carga de trabalho:

A carga de trabalho de cada servidor é determinada como no modelo Hipercubo original de LARSON (1974), aplicado a sistemas sem fila de espera, ou seja:

$$\rho_1 = P_{100} + P_{101} + P_{110} + P_{111} \qquad (3.37)$$
$$\rho_2 = P_{010} + P_{011} + P_{110} + P_{111}$$
$$\rho_3 = P_{001} + P_{011} + P_{101} + P_{111}$$

Note que estas equações diferenciam-se da equação (3.13) do exemplo 1 por não considerar probabilidade de fila $P_q$.

Probabilidade de Perda:

Como o sistema não admite filas, uma chamada de qualquer tipo que chega no sistema quando todos os servidores estão ocupados é considerada perdida. Portanto no exemplo 2, a probabilidade de perda $P_p = P_{111}$.

Freqüências de despacho:

No modelo de múltiplo despacho temos novas medidas de freqüência, tais como:

$f_{ji}^{[1]}$ = fração total de despachos para atendimento de chamadas do tipo 1, nos quais o servidor $j$ é enviado ao átomo $i$;

$f_{(j,k)i}^{[2]}$ = fração total de despachos para atendimento de chamadas tipo 2, nos quais os servidores $j$ e $k$ são enviados simultaneamente ao átomo $i$;

$f_{ji}^{[2]}$ = fração total de despachos para atendimento de chamadas tipo 2, nos quais somente servidor $j$ é enviado ao átomo $i$;

$f_{ji}$ = fração total de despachos (tipo 1 e tipo 2), nos quais o servidor $j$ é enviado ao átomo $i$.

Seja $E_{ji}$ o conjunto de estados nos quais o servidor $j$ é o primeiro servidor disponível na lista de despacho do átomo $i$, ou seja, estados em que o servidor $j$ atende o átomo $i$, $E_{(j,k)i}$ o conjunto de estados nos quais os servidores $j$ e $k$ são os dois primeiros servidores disponíveis da lista de despacho do átomo $i$ (p.e., no exemplo 2: $E_{(1,2)1} = (\{000\},\{001\})$e $B_j$ o estado em que somente o servidor $j$ está disponível no sistema (p.e., no exemplo 2: $B_1$={011}), temos:

(i) Atendimentos tipo 1:

$$f_{ji}^{[1]} = \frac{\frac{\lambda_i^{[1]}}{\lambda^{[1]}} \sum_{B \in E_{ji}} P_B}{(1 - P_{\{111\}})} \tag{3.38}$$

Note que, esta equação corresponde à equação de freqüência do modelo Hipercubo original para sistemas sem fila de espera, já que a mesma só considera chamadas do tipo 1. Desta forma, $\sum_{j}^{N} \sum_{i}^{N_A} f_{ji}^{[1]} = 1$ (3.39)

(ii) Atendimentos tipo 2 :

$$f_{(j,k)i}^{[2]} = \frac{\frac{\lambda_i^{[2]}}{\lambda^{[2]}} \sum_{B \in E_{(j,k)i}} P_B}{(1 - P_{\{111\}})}, \tag{3.40}$$

$$f_{ji}^{[2]} = \frac{\frac{\lambda_i^{[2]}}{\lambda^{[2]}} P_{B_j}}{(1 - P_{\{111\}})}, \tag{3.41}$$

Note, $\sum_{i=1}^{N_A} \left[ \sum_{j=1}^{N} f_{ji}^{[2]} + \sum_{j=1}^{N-1} \sum_{k=j+1}^{N} f_{(j,k)i}^{[2]} \right] = 1$ (3.42)

Podemos definir também outras medidas de freqüências derivadas das apresentadas acima, tais como:

(iii) Atendimento tipo 1:

Freqüência de todos os despachos do sistema que envia o servidor $j$ ao átomo $i$, para atender uma chamada do tipo 1:

$$f'^{[1]}_{ji} = \frac{\frac{\lambda_i^{[1]}}{\lambda} \sum_{B \in E_{ji}} P_B}{(1 - P_{\{111\}})} \tag{3.43}$$

(iv) Atendimento tipo 2:

Freqüência de todos os despachos do sistema que envia o servidor $j$ e $k$ ao átomo $i$, para atender uma chamada do tipo 2:

$$f'^{[2]}_{(j,k)i} = \frac{\frac{\lambda_i^{[2]}}{\lambda} \sum_{B \in E_{(j,k)i}} P_B}{(1 - P_{\{111\}})}, \tag{3.44}$$

Freqüência de todos os despachos do sistema que envia o servidor $j$ ao átomo $i$, para atender uma chamada do tipo 2:

$$f'^{[2]}_{ji} = \frac{\frac{\lambda_i^{[2]}}{\lambda} P_{(B_j)}}{(1 - P_{\{111\}})}, \tag{3.45}$$

Note que,

$$\sum_{i=1}^{N_A} \left[ \sum_{j=1}^{N} (f'^{[1]}_{ji} + f'^{[2]}_{ji}) + \sum_{i=1}^{N-1} \sum_{k=j+1}^{N} f'^{[2]}_{(j,k)i} \right] = 1 \tag{3.46}$$

(v) Atendimentos tipo 1 e tipo 2:

$$f_{ji} = f'^{[1]}_{ji} + f'^{[2]}_{ji} + \sum_{k=j+1}^{N} f'^{[2]}_{(j,k)i} \tag{3.47}$$

Portanto: $$\sum_{j=1}^{N} \sum_{i=1}^{N_A} f_{ji} = 1 \tag{3.48}$$

Tempos de viagem

Os tempos de viagem relacionados aos despachos para atendimento de chamadas tipo 1 são determinados da mesma forma que no modelo Hipercubo para sistemas sem fila de espera, detalhadas em LARSON & ODONI (1981). As expressões de tempo de viagem para

chamadas tipo 1 e tipo 2 são apresentadas a seguir, de forma similar a descrição de CHELST & BARLACH (1981).

(i) Tempo médio de viagem no sistema para chamadas do tipo 1:

A expressão é similar à 3.26, porém não há tempo de espera em fila.

$$\overline{T}^{[1]} = \sum_{j=1}^{N}\sum_{i=1}^{N_A} f_{ji}^{[1]}.t_{ji} \tag{3.49}$$

(ii) Tempo médio de viagem no sistema para chamadas do tipo 2 (considerando que o atendimento inicia-se com a chegada de um dos veículos no local):

$$\overline{T}^{[2]} = \sum_{i=1}^{N_A}\left[\left[\sum_{j=1}^{N-1}\sum_{k=j+1}^{N} f_{(j,k)i}^{[2]} \min(t_{ji}, t_{ki})\right] + \sum_{j=1}^{N} f_{ji}^{[2]} t_{ji}\right] \tag{3.50}$$

(iii) Tempo médio de viagem no sistema para chamadas do tipo 2 (incluindo todos os servidores despachados).

$$\overline{T}_t^{[2]} = \sum_{i=1}^{N_A}\left[\left[\sum_{j=1}^{N-1}\sum_{k=j+1}^{N} f_{(j,k)i}^{[2]} (t_{ji} + t_{ki})\right] + \sum_{j=1}^{N} f_{ji}^{[2]} t_{ji}\right] \tag{3.51}$$

Na equação (3.51) temos que, para cada átomo $i$, o termo $\sum_{j=1}^{N-1}\sum_{k=j+1}^{N} f_{(j,k)i}^{[2]} (t_{ji} + t_{ki})$ refere-se aos atendimentos realizados pelos dois servidores, $j$ e $k$. Note que os tempos dos dois servidores são somados, contabilizando o tempo total de viagem necessário, de ambos os veículos, para atender um único chamado. O termo $\sum_{j=1}^{N} f_{ji}^{[2]} t_{ji}$ refere-se aos atendimentos de chamadas tipo 2 no átomo $i$, nos quais apenas o servidor $j$ atende, pois é o único servidor disponível no sistema.

(iv) Tempo de viagem para o servidor preferencial de chamadas tipo 2 no átomo $i$:

$$\sum_{i=1}^{N_A}\sum_{j=1}^{N}\left[f_{ji}^{[2]} + \sum_{k \in L_{ji}} f_{(j,k)i}^{[2]}\right] t_{ji} \tag{3.52}$$

Na equação, $L'_{ji}$ corresponde ao conjunto de servidores que possuem menor prioridade que o servidor $j$ na lista de despacho do átomo $i$ (p.e, no exemplo 2 (tabela 3.2): $L'_{23}$=(1,3)). Esta expressão determina o tempo médio de viagem para servidor $j$, o qual é a primeira opção da lista de preferência em um atendimento múltiplo despacho. Note que, a expressão considera também chamadas tipo 2, para as quais o servidor $j$ é o único despachado.

Lembre-se que, no caso de patrulha policial, os servidores podem estar em movimento quando disponíveis, e o servidor preferencial pode não ser o primeiro servidor a chegar no local da chamada.

(v) Tempo de viagem para o servidor *backup* de chamadas tipo 2 no átomo $i$ (quando dois servidores $j$ e $k$ são despachados, o servidor preferencial é $j$):

$$\sum_{i=1}^{N_A}\sum_{j=1}^{N}\sum_{k\in L'_{ji}} f_{(j,k)i}^{[2]} t_{ki} \tag{3.53}$$

Esta equação, mede apenas o tempo médio de viagem do ponto de vista do servidor $k$ que participa de um múltiplo despacho como a segunda, terceira,..., opção da lista de despacho.

(vi) Tempo médio de viagem para o primeiro servidor a chegar no local de uma chamada tipo 2:

$$\overline{T}_F = \frac{\sum_{i=1}^{N_A}\sum_{j=1}^{N-1}\sum_{k=j+1}^{N}\sum_{r\in p[j]}\sum_{s\in p[k]} f_{(j,k)i}^{[2]} q_{jr} q_{ks} \min[\tau_{ri},\tau_{si}]}{\sum_{i=1}^{N_A}\sum_{j=1}^{N-1}\sum_{j=k+1}^{N} f_{(j,k)i}^{[2]}} \tag{3.54}$$

(vii) Tempo médio de viagem para o segundo servidor a chegar no local de uma chamada tipo 2:

$$\overline{T}_S = \frac{\sum_{i=1}^{N_A}\sum_{j=1}^{N-1}\sum_{k=j+1}^{N}\sum_{r\in p[j]}\sum_{s\in p[k]} f_{(j,k)i}^{[2]} q_{jr} q_{ks} \max[\tau_{ri},\tau_{si}]}{\sum_{i=1}^{N_A}\sum_{j=1}^{N-1}\sum_{j=k+1}^{N} f_{(j,k)i}^{[2]}} \tag{3.55}$$

Como mencionado anteriormente, as duas últimas medidas de tempo de viagem são mais interessantes para aos sistemas em que o servidor preferencial pode não ser o primeiro a chegar no local da chamada, como no caso de patrulha policial.

Nas equações (3.54) e (3.56), o termo $p[j]$ e $p[k]$ correspondem as áreas nas quais estão localizados os servidores $j$ e $k$, respectivamente. No caso de patrulha policial, os servidores estão se movendo nesta área. Por exemplo, na expressão $r \in p[j]$ significa que o átomo *r* está na área de patrulha do servidor $j$. O termo $q_{jr}$ corresponde a probabilidade do servidor $j$ estar no átomo $r$ e $q_{ks}$ corresponde a probabilidade do servidor $k$ estar no átomo $s$. Os termos $\tau_{ri}$ e $\tau_{si}$ correspondem ao tempo de viagem dos átomos $r$ e $s$, respectivamente, ao átomo $i$. Por exemplo, se uma viatura policial $j$ que está em $r$ e outra viatura *k* que está em $s$, são enviadas simultaneamente para atender uma chamada em $i$, sendo $\tau_{ri} < \tau_{si}$, então $\overline{TF}$ mede o tempo médio para chegada da viatura $j$ e $\overline{TS}$ mede o tempo médio para a chegada da viatura $k$.

Método Aproximado:

O modelo aproximado de CHELST & BARLACH (1981) é similar ao modelo Hipercubo aproximado de LARSON (1975). O método pode ser útil para análise de sistemas de duplo despacho quando N é relativamente grande, pois somente N equações são resolvidas, ao invés de $2^N$. Como no modelo de aproximado de Larson, o método também admite que o estado de um servidor não depende do estado dos demais servidores, e fatores de correção são introduzidos para "corrigir" esta simplificação. Na determinação dos fatores de correção, admite-se que os servidores são homogêneos, ou seja, possuem mesma taxa de atendimento ($\mu_1 = \mu_2 = ... = \mu_N$). A principal diferença do modelo aproximado de CHELST & BARLACH (1981) para o modelo aproximado de LARSON (1975) é que um novo fator de correção é introduzido referente aos atendimentos duplo despacho.

Como o método aproximado não é aplicado no presente estudo, procuramos concentrar as discussões no método exato, mais detalhes podem ser encontrados em CHELST & BARLACH (1981).

# 4. Extensões do modelo Hipercubo para análise dos SAEs em rodovias:

Neste capítulo, discutimos inicialmente as adaptações do modelo Hipercubo, para sua aplicação na análise de SAEs em rodovias com política de único despacho. Algumas destas adaptações já foram estudadas por MENDONÇA & MORABITO (2000, 2001) para avaliar o sistema *Anjos do Asfalto*. A seguir, propomos modificações adicionais necessárias no modelo Hipercubo múltiplo despacho para aplicá-lo na análise de SAEs em rodovias, sendo que estas extensões não estão presentes na literatura. O modelo é então utilizado para representar o SAE da concessionária *Centrovias.*

Uma das principais características dos SAEs em rodovias brasileiras é não admitir fila de espera de chamadas de emergência. Nestes sistemas, quando uma chamada ocorre e os servidores estão ocupados, a mesma é transferida a outro sistema capaz de prestar o serviço requerido, e a chamada é considerada como perda para o sistema. No caso de rodovias, a chamada pode ser transferida, por exemplo, para um outro SAE vizinho ou serviço resgate do corpo de bombeiros da cidade mais próxima. Considerando que a probabilidade de fila $P_q$ é nula, a formulação do modelo Hipercubo para sistemas com fila de capacidade infinita, apresentada no capítulo anterior, deve ser modificada.

Outra importante particularidade dos SAEs em rodovias é a política de despacho. Nestes sistemas, as chamadas não podem ser atendidas por qualquer servidor do sistema, pois devido a longas distâncias dos servidores aos átomos fora de sua área primária, somente alguns servidores podem prestar assistência dentro de um tempo de resposta aceitável para os usuários do sistema. Note que, esta particular política de despacho não respeita uma das hipóteses do modelo Hipercubo original de LARSON (1974), a qual assume que qualquer servidor pode ser despachado a qualquer átomo. Esta específica política de despacho também requer modificações na determinação das equações de equilíbrio dos estados do sistema.

Em alguns SAEs em rodovias, a política de despacho é caracterizada por múltiplo despacho. Nestes casos, uma porcentagem das chamadas do sistema requer o despacho de dois servidores simultaneamente. Quando ocorre um acidente na rodovia, o número de vítimas envolvido pode ser além da capacidade de um servidor. Além disso, os servidores podem ser equipados de forma diferenciada e necessitam trabalhar em conjunto para determinados tipos de acidente, por exemplo, acidentes que envolvem quebra de ferragens ou equipamentos de

controle de incêndio. Como destacado em CHELST & BARLACH (1981), outras interessantes medidas de desempenho podem ser obtidas com relação a chamadas e servidores envolvidos no atendimento tipo múltiplo despacho.

Para análise dos SAEs em rodovias, as nove hipóteses do modelo Hipercubo original apresentadas no capítulo 3 (note as mudanças nas hipóteses 2, 3, 6 e 7) tornam-se:

1. Átomos geográficos: a rodovia é particionada em $N_A$ átomos geográficos, e cada átomo corresponde a uma fonte independente de chamadas.

2. Processo de Chegadas: chamadas de emergência em cada átomo são geradas de acordo com o processo de Poisson de forma independente dos demais átomos com taxa $\lambda_i$. No caso de SAEs com chamadas do tipo 1 (único despacho) e tipo 2 (múltiplo despacho), estas taxas são $\lambda_i^{[1]}$ e $\lambda_i^{[2]}$, respectivamente.

3. Servidores: há N servidores espacialmente distribuídos que podem atender a somente átomos de sua área primária e área *backup*. Denominamos esta situação de *backup* parcial, ou seja, um servidor não pode ser *backup* de qualquer outro servidor (como no modelo Hipercubo original do capítulo 3)

4. Tempo de viagem: o tempo de viagem entre cada par de átomos é conhecido ou pode ser estimado através de conceitos de probabilidade geométrica;

5. Localização dos servidores: cada servidor, quando livre, fica estacionado em uma das bases localizadas ao longo da rodovia;

6. Política de despacho: como mencionado anteriormente, a política de despacho dos SAEs em rodovias é particular, pois somente alguns servidores podem viajar a um determinado átomo (*backup* parcial). Além disso, os SAEs podem operar com políticas de único e múltiplo despacho (quando dois servidores são enviados para atender um chamado do tipo 2).

7. Lista de preferência de despacho: o despacho dos servidores é realizado de acordo com uma lista de preferência. Quando ocorre uma chamada em um determinado átomo, somente os

servidores desta lista podem ser despachados. No caso de único despacho, se o primeiro servidor desta lista estiver disponível, o mesmo é despachado, caso contrário, o próximo servidor disponível da lista é despachado (chamado de servidor *backup*), e assim por diante até o último servidor da lista daquele átomo. No caso de múltiplo despacho, os dois primeiros servidores da lista de preferência de despacho são despachados simultaneamente e se um deles está ocupado, o próximo da lista de despacho (se o número de servidores na lista for maior que dois) é enviado com um dos primeiros que está livre. Se há apenas um servidor disponível na lista de preferência do átomo, o mesmo é despachado sozinho. A lista de preferência é pré-determinada para cada átomo e permanece fixa durante toda a operação do sistema.

8. Tempo de atendimento: o tempo médio de atendimento para cada servidor é conhecido e inclui o tempo de *set-up* (preparação), o tempo de viagem do servidor ao local da chamada, o tempo em cena e o tempo de retorno do servidor a sua base.

9. Tempo de atendimento dependente do tempo de viagem: variações no tempo de atendimento devido a variações no tempo de viagem são consideradas de segunda ordem quando comparadas a variações no tempo em cena ou no tempo de *set-up*.

## 4.1 Adaptação do modelo Hipercubo único despacho para análise dos SAEs em rodovias:

Aqui pretendemos mostrar como adaptar o modelo Hipercubo para análise de um SAE em rodovia no qual apenas um servidor é enviado para atender um chamado. Este exemplo é similar ao sistema *Anjos do Asfalto*, estudado por MENDONÇA & MORABITO (2000, 2001) e descrito brevemente na seção 2.3 do capítulo 2.

**Exemplo ilustrativo 3: SAE em rodovia com único despacho:**

Considere um exemplo de sistema similar ao do exemplo 1 do capítulo 3, com $N = 3$ servidores e $N_A = 4$ átomos, correspondendo a um SAE em um trecho com topologia linear, ilustrado na figura 4.1. Os triângulos representam as bases onde os servidores estão localizados. Admite-se que o sistema não admite filas de espera (capacidade nula), diferente do exemplo 1.

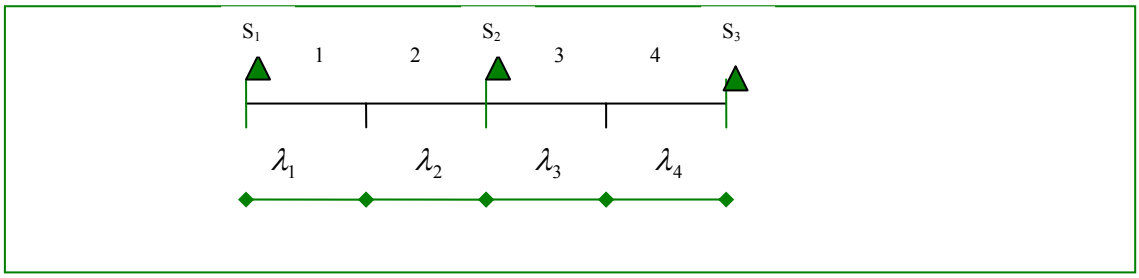


Figura 4.1: SAE em rodovia c/ único despacho - Exemplo 3 (com $N_A$= 4 átomos e N = 3 servidores)

Considerando a política particular de despacho dos SAEs, temos que, para este exemplo, somente dois servidores (ambulâncias) podem ser enviados para atender um despacho (servidor primário ou servidor *backup*). Se a ambulância *backup* estiver ocupada, a chamada é atendida por outro sistema sem ocorrer espera em fila. A lista de preferência de despacho é apresentada na tabela 4.1.

Tabela 4.1 – Lista de preferência de despacho do SAE do exemplo 3(*backup* parcial)

| Átomo | Servidor primário | Servidor *backup* |
|---|---|---|
| 1 | 1 | 2 |
| 2 | 2 | 1 |
| 3 | 2 | 3 |
| 4 | 3 | 2 |

**Equações de equilíbrio:**

Considerando que não há filas, temos $2^3$=8 possíveis estados do sistema: {000}, {001}, {010}, {011}, {100}, {101}, {110}, {111}, como no exemplo 2 do capítulo 3. A seguir apresentamos uma breve discussão das equações de transição entre os estados do sistema.

(iv)Quando todos os servidores estão disponíveis: {000}

No primeiro caso, para o estado em que todos os servidores estão disponíveis, a equação do estado {000} é a mesma do modelo Hipercubo original para fila com capacidade infinita, (equação (3.1) do capítulo 3):

$$(\lambda).P_{\{000\}} = \mu_1.P_{\{100\}} + \mu_2.P_{\{010\}} + \mu_3.P_{\{001\}} \quad (4.1)$$

(v) Quando 1 servidor está ocupado:

A equação também é a mesma para os estados com apenas um servidor ocupado: {001}, {010} e {100} do exemplo 1 no capítulo 3 (equações 3.2 e 3.6). Por exemplo, para o estado {100}:

$$(\lambda + \mu_1).P_{\{100\}} = \lambda_1.P_{\{000\}} + \mu_3.P_{\{101\}} + \mu_2.P_{\{110\}} \quad (4.2)$$

(vi) Quando 2 servidores estão ocupados:

As equações para os estados com dois ou mais servidores ocupados devem ser modificadas, dada a política de despacho de nunca enviar a terceira ambulância mais próxima como *backup*. Por exemplo, para o estado {110}, no qual os servidores 1 e 2 estão ocupados e o servidor 3 está livre, temos:

$$(\lambda_3 + \lambda_4 + \mu_1 + \mu_2).P_{\{110\}} = (\lambda_1 + \lambda_2).P_{\{010\}} + (\lambda_1 + \lambda_2 + \lambda_3).P_{\{100\}} + \mu_3.P_{\{111\}} \quad (4.3)$$

Lado esquerdo da equação (4.3) – fluxo para fora do estado {110}: Analisando as transições para fora do estado {110}, temos:

1. $\{110\} \rightarrow \{111\}$ – ocorre com a chegada de uma chamada nos átomos 3 (taxa $\lambda_3$) ou 4 (taxa $\lambda_4$). Note que, chamadas geradas nos demais átomos não são atendidas pelo sistema;
2. $\{110\} \rightarrow \{010\}$ – ocorre com o término de serviço do servidor 1 (taxa $\mu_1$);
3. $\{110\} \rightarrow \{100\}$ – ocorre com o término de serviço do servidor 2 (taxa $\mu_2$);

Lado direito da equação (4.3) – fluxo para dentro do estado {110}:

1. $\{111\} \rightarrow \{110\}$ - ocorre com o término de serviço do servidor 3 (taxa $\mu_3$);
2. $\{010\} \rightarrow \{110\}$ - ocorre com a chegada de uma chamada no átomo 1 (taxa $\lambda_1$) ou no átomo 2 (taxa $\lambda_2$ - atendimento *backup* );
3. $\{100\} \rightarrow \{110\}$ – ocorre com a chegada de uma chamada nos átomos 1 (taxa $\lambda_1$ - atendimento *backup* ), 2 (taxa $\lambda_2$) e 3 (taxa $\lambda_3$).

(iv) Quando todos os 3 servidores estão ocupados, estado {111}

Com relação ao estado {111} com os três servidores ocupados temos que a equação de equilíbrio torna-se:

$$\mu.P_{\{111\}} = (\lambda_1 + \lambda_2).P_{\{011\}} + \lambda.P_{\{101\}} + (\lambda_3 + \lambda_4).P_{\{110\}} \qquad (4.4)$$

Lado esquerdo da equação (4.4) – fluxo para fora do estado {111}:

Na equação 4.4, as transições para fora de {111} ocorrem com o término de serviço de qualquer servidor. Note que, como não há fila, a transição $\{111\} \rightarrow \{S_4\}$ não ocorre, e as chamadas que chegam quando o sistema está em {111} são perdidas.

Lado direito da equação (4.4) – fluxo para dentro do estado {111}: As transições para dentro de {111}, são:

5. $\{011\} \rightarrow \{111\}$ – ocorre com a chegada de uma chamada no átomo 1 (taxa $\lambda_1$) ou no átomo 2 (taxa $\lambda_2$ - atendimento *backup* );
6. $\{101\} \rightarrow \{111\}$ – ocorre com a chegada de uma chamada em qualquer átomo (taxa total $\lambda$);
7. $\{111\} \rightarrow \{110\}$ – ocorre com a chegada de uma chamada no átomo 3 (taxa $\lambda_3$ - atendimento *backup*) ou no átomo 4 (taxa $\lambda_4$);

As equações de equilíbrio para todos os oito possíveis estados do sistema são:

$$\{000\}\ \lambda.P_{\{000\}} = \mu_3.P_{\{001\}} + \mu_2.P_{\{010\}} + \mu_1.P_{\{100\}} \quad (4.5)$$
$$\{001\}\ (\lambda + \mu_3).P_{\{001\}} = \lambda_4.P_{\{000\}} + \mu_2.P_{\{011\}} + \mu_1.P_{\{101\}}$$
$$\{010\}\ (\lambda + \mu_2).P_{\{010\}} = (\lambda_2 + \lambda_3).P_{\{000\}} + \mu_3.P_{\{011\}} + \mu_1.P_{\{110\}}$$
$$\{011\}\ (\lambda_1 + \lambda_2 + \mu_2 + \mu_3).P_{\{011\}} = (\lambda_2 + \lambda_3 + \lambda_4).P_{\{001\}} + (\lambda_3 + \lambda_4).P_{\{010\}} + \mu_1.P_{\{111\}}$$
$$\{100\}\ \ (\lambda + \mu_1).P_{\{100\}} = \lambda_1.P_{\{000\}} + \mu_2.P_{\{110\}} + \mu_3.P_{\{101\}}$$
$$\{101\}\ \ (\lambda + \mu_1 + \mu_3).P_{\{101\}} = (\lambda_1).P_{\{001\}} + \lambda_4.P_{\{100\}} + \mu_2.P_{\{111\}}$$
$$\{110\}\ \ (\lambda_3 + \lambda_4 + \mu_1 + \mu_2).P_{\{110\}} = (\lambda_1 + \lambda_2).P_{\{010\}} + (\lambda_1 + \lambda_2 + \lambda_3).P_{\{100\}} + \mu_3.P_{\{111\}}$$
$$\{111\}\ \mu.P_{\{111\}} = (\lambda_1 + \lambda_2).P_{\{011\}} + \lambda.P_{\{101\}} + (\lambda_3 + \lambda_4).P_{\{110\}}$$

Para que o sistema acima se torne determinado, podemos substituir uma das equações acima pela equação de normalização. Como a probabilidade de fila é nula e os possíveis estados do sistema são somente {000}, {001}, {010}, {011}, {100}, {101}, {110}, {111}, temos que a equação de normalização é:

$$P_{000} + P_{001} + P_{010} + ... + P_{111} = 1 \quad (4.6)$$

A seguir resolvemos o modelo Hipercubo para o sistema do exemplo 3.

Considere, por exemplo, que as taxas de chegadas nos 4 átomos do sistema são idênticas, $\lambda_i = 0{,}25$ chamadas/unidade de tempo, para $i = 1...4$, Considere também que as taxas de atendimento para os 3 servidores também sejam idênticas, $\mu_j = 1{,}0$ chamadas/unidade de tempo, para $j = 1..3$. Ou seja, $\lambda = 1{,}0$ e $\mu = 3{,}0$. Podemos resolver o sistema acima utilizando, por exemplo, o método de Gauss-Jordan. Os valores encontrados para as probabilidades de estado do sistema são apresentados na tabela 4.2.

Tabela 4.2 - Probabilidades dos estados do sistema

| Estado do sistema | Probabilidade |
|---|---|
| 000 | 0,3852 |
| 001 | 0,1037 |
| 010 | 0,1777 |
| 011 | 0,0815 |
| 100 | 0,1037 |
| 101 | 0,0296 |
| 110 | 0,0815 |
| 111 | 0,0370 |

**Medidas de Desempenho:**

Carga de trabalho dos servidores:

A carga de trabalho de cada servidor $\rho_j$ é calculada considerando todos os estados em que o mesmo se encontra ocupado. Para o exemplo 3 com três servidores temos:

$$\begin{aligned}\rho_1 &= P_{100} + P_{101} + P_{110} + P_{111} \\ \rho_2 &= P_{010} + P_{011} + P_{110} + P_{111} \\ \rho_3 &= P_{001} + P_{011} + P_{101} + P_{111}\end{aligned} \tag{4.7}$$

Os resultados são: $\rho_1 = 0{,}252$, $\rho_3 = 0{,}377$ e $\rho_3 = 0{,}252$

Probabilidade de Perda

No caso de sistemas que não admitem fila e perdem as chamadas quando todos os servidores estão ocupados, uma medida de desempenho interessante é a fração de chamadas perdidas do sistema. No caso dos SAEs em rodovias, em particular, o sistema não perde apenas chamadas no estado em que todos os servidores estão ocupados. Como somente determinados servidores do sistema podem viajar a um dado átomo para atender uma chamada, uma perda também pode ocorrer em outros estados do sistema, mesmo com servidores disponíveis. Para o exemplo 3 acima, a probabilidade de perda $P_p$ pode ser calculada da seguinte forma:

$$P_p = \frac{(\lambda_1 + \lambda_2)}{\lambda} P_{110} + \frac{(\lambda_3 + \lambda_4)}{\lambda} P_{011} + P_{111} \tag{4.8}$$

Observe que cada termo da soma acima corresponde ao produto da probabilidade de uma chamada chegar no átomo $i$, pela soma das probabilidades dos estados do sistema em que os servidor preferencial e o servidor *backup* do átomo $i$ estão ocupados. Por exemplo, o termo $\frac{(\lambda_1 + \lambda_2)}{\lambda} P_{110}$, mostra que uma chamada no átomo 1 ou 2 é perdida se o sistema estiver no estado $\{110\}$, pois o servidor 3 não pode atender chamadas nestes átomos, e os dois servidores de sua lista de despacho estão ocupados. O valor de $P_p$, resolvendo a equação acima para o exemplo 3, é 0,1185.

Freqüência de despacho:

(i) Fração de todos os despachos do servidor $j$ ao átomo $i$:

LARSON & ODONI (1981) definiram $f_{ji}$ (modelo Hipercubo original ) para sistemas que não admitem fila. No caso de um sistema com N = 3 servidores, esta expressão é dada por:

$$f_{ji} = f_{ji}^{[nq]} = \frac{\frac{\lambda_i}{\lambda} \sum_{B \in E_{ji}} P_B}{(1 - P_{111})} \tag{4.9}$$

O termo $f_{ji}^{[q]}$, presente no cálculo de $f_{ji}$ para sistemas com fila de capacidade infinita (equação (3.16) do capítulo 3), não aparece na expressão acima, pois não há espera de chamadas em fila. Além disso, é preciso considerar a probabilidade condicionada de que a chamada atendida não espera em fila. Por isso dividimos $f_{ji}^{[nq]}$ da equação 3.16 pelo termo $(1 - P_{111})$, como na expressão (4.9).

No caso do SAE em rodovia representado no exemplo 3, $f_{ji}$ deve ser calculado por:

$$f_{ji} = f_{ji}^{[nq]} = \frac{\frac{\lambda_i}{\lambda} \sum_{B \in E_{ji}} P_B}{(1 - P_p)} \tag{4.10}$$

Observe que, ao invés do termo $(1 - P_{111})$, temos $(1 - P_p)$, pois a chamada não é somente perdida quando todos os servidores estão ocupados, mas também quando os dois possíveis servidores de um átomo de onde provém a chamada estiverem ocupados. Por exemplo, para o cálculo de $f_{12}$ temos:

$$f_{22} = f_{22}^{[nq]} = \frac{\frac{\lambda_2}{\lambda}\left(P_{\{000\}} + P_{\{100\}} + P_{\{001\}} + P_{\{101\}}\right)}{(1 - P_p)} \tag{4.11}$$

Observe que, $E_{22} = (\{000\}, \{100\}, \{001\}, \{101\})$, pois nestes estados o servidor 2 está livre e chamadas no átomo 2 devem ser atendidas pelo servidor preferencial 2. A tabela 4.3 apresenta as freqüências de despachos do servidor $j$ ao átomo $i$ no sistema do exemplo 3. Note que,

$$\sum_{j}^{N} \sum_{i}^{N_A} f_{ji} = 1$$

Tabela 4.3 – Freqüências de despacho do servidor $j$ ao átomo $i$ ( $f_{ji}$ )

| $f_{ji} = f_{ji}^{[nq]}$ | **1** | **2** | **3** | **4** |
|---|---|---|---|---|
| 1 | 0,2121 | 0,0735 | 0 | 0 |
| 2 | 0,0378 | 0,1764 | 0,1764 | 0,0378 |
| 3 | 0 | 0 | 0,0735 | 0,2121 |

As expressões referentes à freqüência de atendimento *backup* são obtidas da mesma forma que para os sistemas com fila de capacidade infinita.

Tempos de viagem :

Os expressões para os tempos de viagem devem considerar $P_q = 0$ e $T_q = 0$, pois o sistema não admite filas. Para calcular o tempo de viagem do servidor $j$ ao átomo $i$, $t_{ji}$, devemos considerar que os servidores no SAE estão fixos em suas bases. No exemplo 3 acima, a matriz dos tempos de viagem do servidor $j$ ao átomo $i$, pode ser calculada por meio da distância dos servidores ao centróide do átomo em que o mesmo viaja. Supomos que os valores de $t_{ji}$ em minutos são dados na tabela 4.4:

Tabela 4.4 – Tempo de viagem servidor $j$ - átomo $i$ ( $t_{ji}$ )

| $t_{ji}$ (min) | **1** | **2** | **3** | **4** |
|---|---|---|---|---|
| 1 | 5,0 | 8,0 | 10,0 | 15,0 |
| 2 | 8,0 | 5,0 | 5,0 | 10,0 |
| 3 | 15,0 | 10,0 | 8,0 | 5,0 |

As expressões das mais importantes medidas relacionadas ao tempo de viagem para o sistema são apresentadas a seguir:

(i) Tempo médio de viagem no sistema:

$$\overline{T} = \sum_{j=1}^{N} \sum_{i=1}^{N_A} f_{ji} . t_{ji} \qquad (4.12)$$

No exemplo 3 com N = 3 servidores, $\overline{T} = 5{,}744$ minutos

(ii) Tempo médio de viagem ao átomo $i$ :

$$\overline{T}_i = \frac{\sum_{j=1}^{N} f_{ji} . t_{ji}}{\sum_{j=1}^{N} f_{ji}} \tag{4.13}$$

(iii) Tempo médio de viagem de cada servidor $j$ :

$$\overline{TU}_j = \frac{\sum_{i=1}^{N_A} f_{ji} . t_{ji}}{\sum_{i=1}^{N_A} f_{ji}} \tag{4.14}$$

Os valores de $\overline{TU}_j$ para $j = 1,2,3$ no sistema do exemplo 3 são: 5,77 min, 5,70 min e 5,77 min, respectivamente.

O modelo Hipercubo apresentado nesta seção foi utilizado para análise do sistema *Anjos do Asfalto*, e o resultados desta aplicação são apresentados no capítulo 6.

## 4.2 Adaptação do modelo Hipercubo múltiplo despacho para análise dos SAEs em rodovias:

O modelo proposto por CHELST & BARLACH (1981) e revisado no capítulo 3 pode ser adaptado para análise dos SAEs em rodovias, considerando que alguns destes sistemas são caracterizados por:

- Duplo despacho de servidores para atender um mesmo acidente (chamadas do tipo 2), que podem ser servidores idênticos (dois resgates) ou diferentes (carro médico e resgate);

- Não há fila de espera e a probabilidade de perda é uma medida importante, sendo que uma chamada pode ser perdida mesmo se há servidores disponíveis;

- A política de despacho deve considerar que somente alguns servidores do sistema podem atender um determinado átomo (*backup* parcial).

Apresentamos a seguir como o sistema em rodovia do exemplo 3, pode ser analisado considerando múltiplo despacho.

**Exemplo ilustrativo 4**: sistema SAE em rodovia c/ múltiplo despacho

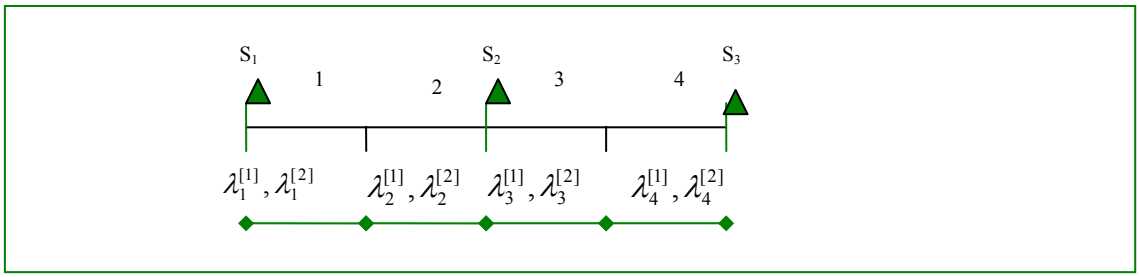


Figura 4.2 : SAE em rodovia - múltiplo despacho - Exemplo 4 (com $N_A$= 4 átomos e N = 3 servidores)

O sistema da figura 4.2 possui as mesmas características do sistema do exemplo 3, com relação ao número e disposição dos servidores ($N$ = 3 servidores), distribuição dos átomos ($N_A$ = 4 átomos) e lista de preferência de despacho para cada átomo (tabela 4.1). No entanto, neste sistema, temos que, em cada átomo $i$, pode ocorrer chamadas tipo 2 ($\lambda_i^{[2]}$), que requerem despacho de dois servidores simultaneamente (duplo despacho).

Com relação à política de despacho, uma chamada tipo 1 pode ser atendida por um servidor preferencial ou um *backup* (como no exemplo 3). Uma chamada tipo 2 é atendida por dois servidores ao mesmo tempo, e é perdida pelo sistema se os dois servidores preferenciais estiverem ocupados. Note que, os dois servidores preferenciais de uma chamada tipo 2 de um dado átomo corresponde aos servidores preferencial e *backup* de uma chamada tipo 1 no mesmo átomo. Desta forma, a lista de preferência de despacho do átomo é mesma para chamadas tipo 1 e tipo 2. Além disso, se uma chamada tipo 2 ocorre quando um dos dois servidores estiver ocupado, o outro disponível deve atender sem ajuda de um terceiro servidor do sistema.

**Transição de estados:**

Similarmente aos exemplos 2 e 3 (sistemas sem filas), os $2^3$ = 8 estados possíveis do sistema são os vértices da figura 4.3 : {000}, {001}, {010}, {011}, {100}, {101}, {110}, {111}.

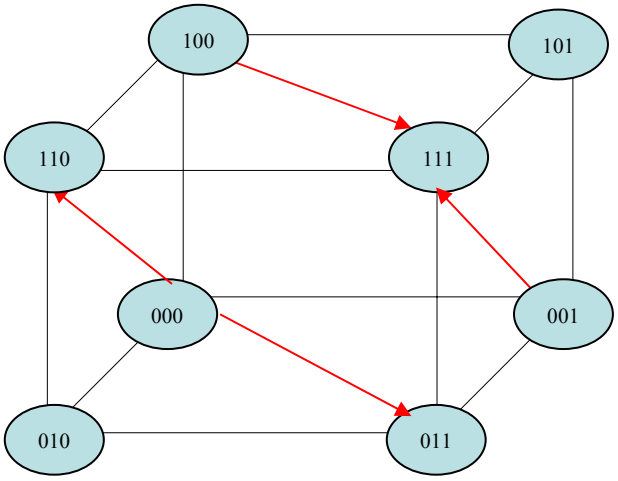


Figura 4.3 – Cubo representando os estados do sistema

Note na figura 4.3 que as transições adicionais resultantes da política de múltiplo despacho são: $\{000\} \rightarrow \{011\}, \{000\} \rightarrow \{110\}, \{100\} \rightarrow \{111\}$ e $\{001\} \rightarrow \{111\}$. Comparando a figura 3.8 (exemplo 2) com a figura 4.3, observamos que a transição diagonal $\{010\} \rightarrow \{111\}$ que ocorre no exemplo 2 não ocorre no exemplo 3. Isso porque, de acordo com a política de despacho particular deste SAE e a lista de preferência de despacho (tabela 4.1), os servidores 1 e 3 não são despachados juntos, e uma chamada tipo 2 que chega no sistema no estado $\{010\}$ é atendida por apenas um servidor.

Além disso, como a política de despacho deste sistema não considera a possibilidade de escalar uma terceira opção na lista de preferência dos átomos do sistema (tabela 4.1), devemos observar que, diferentemente do exemplo 2: a transição $\{001\} \rightarrow \{111\}$ somente ocorre com a chegada de uma chamada tipo 2 nos átomos 1 ou 2 (cujos dois servidores preferenciais são 1 e 2). A transição $\{100\} \rightarrow \{111\}$ somente ocorre com a chegada de uma chamada tipo 2 nos átomos 3 ou 4 (para os quais o servidor 2 e 3 são os dois preferenciais). No exemplo 2 estas transições ocorrem com a chegada de uma chamada em qualquer átomo.

As transições $\{000\} \rightarrow \{011\}$ e $\{000\} \rightarrow \{110\}$ ocorrem de forma similar ao exemplo 2, dado que os dois servidores preferenciais de todos os átomos estão disponíveis no estado $\{000\}$.

**Equações de equilíbrio:**

A seguir, apresentamos uma breve discussão das equações de transição entre os estados do sistema. Como no exemplo 2, admitimos que $\lambda^{[1]} = \sum_{i=1}^{N_A} \lambda_i^{[1]}$ , $\lambda^{[2]} = \sum_{i=1}^{N_A} \lambda_i^{[2]}$ , $\lambda = \lambda^{[1]} + \lambda^{[2]}$ e $\mu = \sum_{j=1}^{N} \mu_j$ .

(i) Quando todos os servidores estão disponíveis: {000}

Neste caso a equação é a mesma do modelo Hipercubo com múltiplo despacho, discutido para ao sistema do exemplo 2 (equação (3.31) do capítulo 3), assim como a equação (4.1) se $\lambda = \lambda^{[1]} + \lambda^{[2]}$.

$$(\lambda).P_{\{000\}} = \mu_1.P_{\{100\}} + \mu_2.P_{\{010\}} + \mu_3.P_{\{001\}} \qquad (4.15)$$

(ii) Quando há apenas 1 servidor ocupado no sistema:

Considerando que $\lambda = \lambda^{[1]} + \lambda^{[2]}$, as equações de equilíbrio dos estados {001}, {010} e {100} no exemplo 4 são similares às equações de equilíbrio destes estados no exemplo 3 (compare com as equações (4.2) e (4.5)), ou seja:

$$\begin{aligned}
&\{001\}\ (\lambda + \mu_3).P_{\{001\}} = \lambda_4^{[1]}.P_{\{000\}} + \mu_2.P_{\{011\}} + \mu_1.P_{\{101\}} \\
&\{010\}\ (\lambda + \mu_2).P_{\{010\}} = (\lambda_2^{[1]} + \lambda_3^{[1]}).P_{\{000\}} + \mu_1.P_{\{110\}} + \mu_3.P_{\{011\}} \\
&\{100\}\ (\lambda + \mu_1).P_{\{100\}} = \lambda_1^{[1]}.P_{\{000\}} + \mu_3.P_{\{101\}} + \mu_2.P_{\{110\}}
\end{aligned} \qquad (4.16)$$

Note que, uma chamada tipo 2 que encontra um dos seus servidores ocupados é atendida unicamente pelo outro servidor livre. Por exemplo, se uma chamada tipo 2 ocorre no átomo 1 quando o sistema está em {100}, a mesma é atendida pelo servidor 2 e ocorre a transição $\{100\} \rightarrow \{110\}$, ao invés de $\{100\} \rightarrow \{111\}$ como no exemplo 2.

(iii) Quando há 2 servidores ocupados no sistema:

Analisando por exemplo o estado {110}, temos a equação de equilíbrio:

$$\begin{aligned}
&(\lambda_3^{[1]} + \lambda_4^{[1]} + \lambda_3^{[2]} + \lambda_4^{[2]} + \mu_1 + \mu_2).P_{\{110\}} = (\lambda_1^{[2]} + \lambda_2^{[2]}).P_{000} + (\lambda_1^{[1]} + \lambda_2^{[1]} + \lambda_1^{[2]} + \lambda_2^{[2]}).P_{\{010\}} + \\
&(\lambda_1^{[1]} + \lambda_2^{[1]} + \lambda_3^{[1]} + \lambda_1^{[2]} + \lambda_2^{[2]}).P_{\{100\}} + \mu_3.P_{\{111\}}
\end{aligned} \qquad (4.17)$$

Lado esquerdo da equação (4.17) – fluxo para fora do estado {110}: Analisando as possíveis transições para fora deste estado, observamos que a equação (4.17) é diferente da equação de equilíbrio deste estado (3.33 ) do capítulo 3, e assim temos:

4. $\{110\} \rightarrow \{111\}$ – ocorre com a chegada de uma chamada do tipo 1 ou tipo 2, nos átomos 3 e 4 (taxa total $\lambda_3^{[1]} + \lambda_4^{[1]} + \lambda_3^{[2]} + \lambda_4^{[2]}$). Lembre-se que no exemplo 2, esta transição ocorre com uma chamada gerada em qualquer átomo do sistema;
5. $\{110\} \rightarrow \{010\}$ – ocorre com o término de serviço do servidor 1 (taxa $\mu_1$);
6. $\{110\} \rightarrow \{100\}$ – ocorre com o término de serviço do servidor 2 (taxa $\mu_2$).

Lado direito da equação (4.17) – fluxo para dentro do estado {110}: As transições para dentro do estado {110} são:

5. $\{111\} \rightarrow \{110\}$ – ocorre com o término de serviço do servidor 3 (taxa $\mu_3$);
6. $\{010\} \rightarrow \{110\}$ – ocorre com a chegada de uma chamada tipo 1 ou tipo 2 no átomo 1 (taxa $\lambda_1^{[1]} + \lambda_1^{[2]}$) ou no átomo 2 (taxa $\lambda_2^{[1]} + \lambda_2^{[2]}$). Note que, diferentemente do exemplo 2, uma chamada tipo 2 é atendida por apenas um servidor disponível da sua lista de despacho (servidor 1);
7. $\{100\} \rightarrow \{110\}$ – ocorre com a chegada de uma chamada do tipo 1 nos átomos 1, 2 ou 3 (taxa total $\lambda_1^{[1]} + \lambda_2^{[1]} + \lambda_3^{[1]}$) e do tipo 2 nos átomos 1 e 2 (taxa total $\lambda_1^{[2]} + \lambda_2^{[2]}$), diferentemente da equação (3.33);
8. $\{000\} \rightarrow \{110\}$ – ocorre com a chegada de uma chamada do tipo 2 nos átomos 1 ou 2 (taxa total $\lambda_1^{[2]} + \lambda_2^{[2]}$ ), como no exemplo 2.

Em particular, para o estado {101}, em que somente o servidor 2 está disponível, a equação de equilíbrio no exemplo 4 é a mesma que no exemplo 2 (veja equações (3.35)), dado que o servidor 2 pode atender qualquer átomo (como preferencial ou *backup*), assim temos: :

$$(\lambda + \mu_1 + \mu_3).P_{\{101\}} = \lambda_1^{[1]}.P_{\{001\}} + \lambda_4^{[1]}.P_{\{100\}} + \mu_2.P_{\{111\}} \qquad (4.18)$$

(iv)Quando todos os servidores estão ocupados: {111}

$$\mu.P_{\{111\}} = (\lambda_1^{[1]} + \lambda_2^{[1]} + \lambda_1^{[2]} + \lambda_2^{[2]}).P_{\{011\}} + \lambda.P_{\{101\}} + (\lambda_3^{[1]} + \lambda_4^{[1]} + \lambda_3^{[2]} + \lambda_4^{[2]}).P_{\{110\}} + (\lambda_1^{[2]} + \lambda_2^{[2]}).P_{\{001\}} + (\lambda_3^{[2]} + \lambda_4^{[2]}).P_{\{100\}} \qquad (4.19)$$

Lado esquerdo da equação (4.19) – fluxo para fora do estado {111}:

Como nos exemplos 2 e 3, o sistema sai do estado {111} com o término de serviço de qualquer servidor, e a taxa total é $\mu$.

Lado direito da equação (4.19) – fluxo para dentro do estado {111}: As transições para dentro do estado {111} diferem das transições do exemplo 2 para este estado (compare a equação (4.19) com a equação (3.34) do capítulo 3). Assim :

1. $\{011\} \rightarrow \{111\}$ – ocorre com a chegada de uma chamada do tipo 1 ou tipo 2 nos átomos 1 ou 2 (taxa total $\lambda_1^{[1]} + \lambda_2^{[1]} + \lambda_1^{[2]} + \lambda_2^{[2]}$);
2. $\{110\} \rightarrow \{111\}$ – ocorre com a chegada de uma chamada do tipo 1 ou tipo 2 nos átomos 3 ou 4 (taxa total $\lambda_3^{[1]} + \lambda_4^{[1]} + \lambda_3^{[2]} + \lambda_4^{[2]}$);
3. $\{101\} \rightarrow \{111\}$ – como no exemplo 2, ocorre com a chegada de uma chamada de qualquer tipo em qualquer átomo. Note na tabela 4.1 que o servidor 2 pode atender todos os átomos como preferencial ou *backup*;
4. $\{001\} \rightarrow \{111\}$- ocorre com a chegada de uma chamada do tipo 2 nos átomos 1 ou 2 (taxa $\lambda_1^{[2]} + \lambda_2^{[2]}$);
5. $\{100\} \rightarrow \{111\}$- ocorre com a chegada de uma chamada do tipo 2 nos átomos 3 ou 4 (taxa $\lambda_3^{[2]} + \lambda_4^{[2]}$).

Como discutido anteriormente, a transição $\{010\} \rightarrow \{111\}$, que ocorre no exemplo 2 não ocorre no exemplo 4.

O sistema de equações de equilíbrio para os oito possíveis estados {000},{001}, {010},{011},{100}, {101},{110} e {111} é :

$$\{000\}\ \lambda.P_{\{000\}} = \mu_1.P_{\{100\}} + \mu_2.P_{\{010\}} + \mu_3.P_{\{001\}} \qquad (4.20)$$

$$\{001\}\ (\lambda + \mu_3).P_{\{001\}} = \lambda_4^{[1]}.P_{\{000\}} + \mu_2.P_{\{011\}} + \mu_1.P_{\{101\}}$$

$$\{010\}\ (\lambda + \mu_2).P_{\{010\}} = (\lambda_2^{[1]} + \lambda_3^{[1]}).P_{\{000\}} + \mu_1.P_{\{110\}} + \mu_3.P_{\{011\}}$$

$$\{011\}\ \begin{aligned}&(\lambda_1^{[1]} + \lambda_2^{[1]} + \lambda_1^{[2]} + \lambda_2^{[2]} + \mu_2 + \mu_3).P_{\{011\}} = (\lambda_3^{[2]} + \lambda_4^{[2]}).P_{000} + (\lambda_3^{[1]} + \lambda_4^{[1]} + \lambda_3^{[2]} + \lambda_4^{[2]}).P_{\{010\}} + \\ &(\lambda_2^{[1]} + \lambda_3^{[1]} + \lambda_4^{[1]} + \lambda_3^{[2]} + \lambda_4^{[2]}).P_{\{001\}} + \mu_1.P_{\{111\}}\end{aligned}$$

$\{100\}\ (\lambda+\mu_1).P_{\{100\}} = \lambda_1^{[1]}.P_{\{000\}} + \mu_3.P_{\{101\}} + \mu_2.P_{\{110\}}$

$\{101\}\ (\lambda+\mu_1+\mu_3).P_{\{101\}} = \lambda_1^{[1]}.P_{\{001\}} + \lambda_4^{[1]}.P_{\{100\}} + \mu_2.P_{\{111\}}$

$\{110\}$

$(\lambda_3^{[1]}+\lambda_4^{[1]}+\lambda_3^{[2]}+\lambda_4^{[2]}+\mu_1+\mu_2).P_{\{110\}} = (\lambda_1^{[2]}+\lambda_2^{[2]}).P_{000} + (\lambda_1^{[1]}+\lambda_2^{[1]}+\lambda_1^{[2]}+\lambda_2^{[2]}).P_{\{010\}} +$

$(\lambda_1^{[1]}+\lambda_2^{[1]}+\lambda_3^{[1]}+\lambda_1^{[2]}+\lambda_2^{[2]}).P_{\{100\}} + \mu_3.P_{\{111\}}$

$\{111\}\ \mu.P_{\{111\}} = (\lambda_1^{[1]}+\lambda_2^{[1]}+\lambda_1^{[2]}+\lambda_2^{[2]}).P_{\{011\}} + (\lambda).P_{\{101\}} + (\lambda_3^{[1]}+\lambda_4^{[1]}+\lambda_3^{[2]}+\lambda_4^{[2]}).P_{\{110\}} +$
$(\lambda_1^{[2]}+\lambda_2^{[2]}).P_{\{001\}} + (\lambda_3^{[2]}+\lambda_4^{[2]}).P_{\{100\}}$

Para determinar algumas medidas de desempenho através do modelo Hipercubo no sistema do exemplo 4, considere idênticas as taxas de chegadas tipo 1 e tipo 2 nos 4 átomos do sistema, com $\lambda_i^{[1]} = 0{,}20$ e $\lambda_i^{[2]} = 0{,}05$ chamadas/unidade de tempo, para $i = 1...4$, ou seja, $\lambda^{[1]} = 0{,}80$, $\lambda^{[2]} = 0{,}20$ e $\lambda = 1{,}0$. Considere também que as taxas de atendimento para os 3 servidores sejam iguais, $\mu_j = 1{,}0$ chamadas/unidade de tempo, para $j = 1..3$, ou seja, $\mu = 3{,}0$. Podemos resolver o sistema acima utilizando, por exemplo, o método de Gauss-Jordan. Os valores encontrados para as probabilidades de estado do sistema são apresentados na tabela 4.5.

Tabela 4.5 - Probabilidades dos estados do sistema

| Estado do sistema | Probabilidade |
|---|---|
| 000 | 0,3661 |
| 001 | 0,0988 |
| 010 | 0,1684 |
| 011 | 0,0952 |
| 100 | 0,0988 |
| 101 | 0,0292 |
| 110 | 0,0952 |
| 111 | 0,0481 |
| soma | 1,0 |

**Medidas de Desempenho:**

Carga de trabalho dos servidores

Como nos exemplos 2 e 3 anteriores, a carga de trabalho de cada servidor $\rho_j$ é calculada considerando todos os estados em que o mesmo se encontra ocupado. Para o exemplo com três servidores temos:

$$\rho_1 = P_{100} + P_{101} + P_{110} + P_{111} \quad (4.21)$$
$$\rho_2 = P_{010} + P_{011} + P_{110} + P_{111}$$
$$\rho_3 = P_{001} + P_{011} + P_{101} + P_{111}$$

Os valores obtidos para o exemplo 4, são: $\rho_1 = 0{,}271, \rho_2 = 0{,}407$ e $\rho_3 = 0{,}271$.

Probabilidade de Perda

De acordo com a discussão no exemplo 3, a probabilidade de perda é uma medida importante para os SAEs, pois uma chamada pode ser perdida mesma quando há servidores disponíveis. No caso de SAEs em rodovias com política de múltiplo despacho há três medidas de probabilidade de perda: probabilidade de perda de chamadas tipo 1 ($P_p^{[1]}$), probabilidade de perda de chamadas tipo 2 ($P_p^{[2]}$) e probabilidade de perda para qualquer chamada do sistema ($P_p$). Para o exemplo 4, estas medidas são calculadas da seguinte forma:

$$P_p^{[1]} = \frac{(\lambda_1^{[1]} + \lambda_2^{[1]})}{\lambda^{[1]}} P_{\{110\}} + \frac{(\lambda_3^{[1]} + \lambda_4^{[1]})}{\lambda^{[1]}} P_{\{011\}} + P_{\{111\}} \quad (4.22)$$

$$P_p^{[2]} = \frac{(\lambda_1^{[2]} + \lambda_2^{[2]})}{\lambda^{[2]}} P_{\{110\}} + \frac{(\lambda_3^{[2]} + \lambda_4^{[2]})}{\lambda^{[1]}} P_{\{011\}} + P_{\{111\}} \quad (4.23)$$

$$P_p = \frac{(\lambda_1^{[1]} + \lambda_2^{[1]} + \lambda_1^{[2]} + \lambda_2^{[2]})}{\lambda} P_{\{110\}} + \frac{(\lambda_3^{[1]} + \lambda_4^{[1]} + \lambda_1^{[2]} + \lambda_2^{[2]})}{\lambda} P_{\{011\}} + P_{\{111\}} \quad (4.24)$$

Observe nas expressões acima que, uma chamada tipo 1 ou tipo 2 não é atendida se seus dois servidores estão ocupados. Por exemplo, no estado 110, qualquer chamada gerada nos átomos 1 ou 2 é perdida para o sistema, pois o servidor 3 não pode atendê-la. Assim, o resultado destas medidas para o exemplo 4 é $P_p^{[1]} = 0{,}1433$ , $P_p^{[2]} = 0{,}1433$ e $P_p = 0{,}1433$. Note que, por coincidência neste exemplo, estas medidas resultam iguais, pois $\left( \frac{\lambda_i^{[1]}}{\lambda^{[1]}} = \frac{\lambda_i^{[2]}}{\lambda^{[2]}} = \frac{\lambda_i^{[2]} + \lambda_i^{[2]}}{\lambda} \right)$.

Freqüências de despacho

(i) Freqüência de despachos do servidor $j$ ao átomo $i$ (considerando tipo 1 e tipo 2 independentes):

Para chamadas tipo 1:

$$f_{ji}^{[1]} = \frac{\frac{\lambda_i^{[1]}}{\lambda^{[1]}} \sum_{B \in E_{ji}} P_B}{(1 - P_p^{[1]})} \tag{4.25}$$

Por exemplo, para o cálculo de $f_{12}^{[1]}$ temos:

$$f_{12}^{[1]} = \frac{\frac{\lambda_2^{[1]}}{\lambda^{[1]}} \left( P_{\{010\}} + P_{\{011\}} \right)}{(1 - P_p^{[1]})} \tag{4.26}$$

Note que, $E_{12} = \left( \{010\}, \{011\} \right)$, pois nestes estados o servidor 2 está ocupado e chamadas no átomo 2 devem ser atendidas pelo servidor *backup* 1. A tabela 4.6 apresenta os valores de $f_{ji}^{[1]}$ para o exemplo 4.

Tabela 4.6 – Freqüências de despacho do servidor $j$ ao átomo $i$ ( $f_{ji}^{[1]}$ )

| $f_{ji}^{[1]}$ | **1** | **2** | **3** | **4** |
|---|---|---|---|---|
| 1 | 0,2126 | 0,0770 | 0 | 0 |
| 2 | 0,0374 | 0,1730 | 0,1730 | 0,0374 |
| 3 | 0 | 0 | 0,0770 | 0,2126 |

Note que como descrito no capítulo 3, expressão (3.39), temos pelos valores da tabela 4.3 que:

$$\sum_{j}^{3} \sum_{i}^{4} f_{ji}^{[1]} = 1 \tag{4.27}$$

Para chamadas do tipo 2 – despacho de 2 servidores:

$$f_{(j,k)i}^{[2]} = \frac{\frac{\lambda_i^{[2]}}{\lambda^{[2]}} \sum_{B \in E_{(j,k)i}} P_B}{(1 - P_p^{[2]})}, \tag{4.28}$$

onde $f^{[2]}_{(j,k)i}$ refere-se a proporção de atendimentos tipo 2 no átomo $i$ realizados pelos servidores $j$ e $k$, e portanto $f^{[2]}_{(j,k)i} = 0$, se $j$ e $k$ não são seus dois servidores preferenciais. Por exemplo, para o cálculo de $f^{[2]}_{(1,2)2}$ no exemplo 4, temos:

$$f^{[2]}_{(1,2)2} = \frac{\frac{\lambda^{[2]}_2}{\lambda^{[2]}}\left(P_{\{000\}} + P_{\{001\}}\right)}{(1 - P^{[2]}_p)} \tag{4.29}$$

Os valores de $f^{[2]}_{(j,k)i}$ para o exemplo 4, considerando que *j* é o servidor preferencial e *k* é o servidor *backup* do átomo *i*, são: $f^{[2]}_{(1,2)1} = f^{[2]}_{(2,1)2} = f^{[2]}_{(2,3)3} = f^{[2]}_{(3,2)4} = 0{,}1357$. Note que, como mencionado anteriormente, neste exemplo em particular, estes valores são iguais pois: $\left(\frac{\lambda^{[1]}_i}{\lambda^{[1]}} = \frac{\lambda^{[2]}_i}{\lambda^{[2]}} = \frac{\lambda^{[2]}_i + \lambda^{[2]}_i}{\lambda}\right)$.

Para chamadas do tipo 2 – despacho de 1 servidor:

$$f^{[2]}_{ji} = \frac{\frac{\lambda^{[2]}_i}{\lambda^{[2]}} \sum_{B \in F_{ji}} P_B}{(1 - P^{[2]}_p)}, \tag{4.30}$$

Diferentemente da expressão (3.41), o termo $F_{ji}$ da expressão (4.30) deve ser definido como um conjunto de estados em que somente o servidor $j$ pode atender uma chamada em $i$ (p.e., no exemplo 4: $F_{22} = (\{100\}, \{101\})$ ). Lembre-se que, devido a política de despacho parcial, há mais de um estado em que o servidor *j* é o único servidor possível para atender uma chamada no átomo *i*, mesmo que hajam outros disponíveis no sistema. Por exemplo, para o cálculo de $f^{[2]}_{22}$ no exemplo 4, temos:

$$f^{[2]}_{22} = \frac{\frac{\lambda^{[2]}_2}{\lambda^{[2]}}\left(P_{\{100\}} + P_{\{101\}}\right)}{(1 - P^{[2]}_p)} \tag{4.31}$$

Observe que se o sistema está no estado {100}, somente o servidor 2 é despachado para atender uma chamada tipo 2 no átomo 2, mesmo que o servidor 3 esteja disponível. A tabela 4.7 apresenta os valores de $f^{[2]}_{ji}$ para o exemplo 4.

Tabela 4.7 – Freqüências de despacho do servidor $j$ ao átomo $i$ ( $f_{ji}^{[2]}$ )

| $f_{ji}^{[2]}$ | **1** | **2** | **3** | **4** |
|---|---|---|---|---|
| 1 | 0,0770 | 0,0770 | 0 | 0 |
| 2 | 0,0374 | 0,0374 | 0,0374 | 0,0374 |
| 3 | 0 | 0 | 0,0770 | 0,0770 |

Temos então que, como na expressão (3.42) do capítulo 3 :

$$\sum_{i=1}^{4}\left[\sum_{j=1}^{3} f_{ji}^{[2]} + \sum_{j=1}^{2}\sum_{k=j+1}^{3} f_{(j,k)i}^{[2]}\right] = 1 \qquad (4.32)$$

(ii) Freqüência de despachos do servidor $j$ ao átomo $i$ (considerando todos os despachos):

Freqüência de todos os despachos do sistema que envia o servidor $j$ ao átomo $i$, para atender uma chamada do tipo 1:

$$f'^{[1]}_{ji} = \frac{\dfrac{\lambda_i^{[1]}}{\lambda}\sum_{B\in E_{ji}} P_B}{(1-P_p)} \qquad (4.33)$$

Freqüência de todos os despachos do sistema que envia o servidor $j$ e $k$ ao átomo $i$, para atender uma chamada do tipo 2:

$$f'^{[2]}_{(j,k)i} = \frac{\dfrac{\lambda_i^{[2]}}{\lambda}\sum_{B\in E_{(j,k)i}} P_B}{(1-P_p)}, \qquad (4.34)$$

Freqüência de todos os despachos do sistema que envia o servidor $j$ ao átomo $i$, para atender uma chamada do tipo 2:

$$f'^{[2]}_{ji} = \frac{\dfrac{\lambda_i^{[2]}}{\lambda}\sum_{B\in F_{ji}} P_B}{(1-P_p)} \qquad (4.35)$$

Note que, como na expressão (3.46) e (3.48), temos:

$$\sum_{j=1}^{N}\sum_{i=1}^{N_A}\left[ f'^{[1]}_{ji} + f'^{[2]}_{ji} + \sum_{l=1}^{N-1} f'^{[2]}_{(j,k)i}\right] = 1 \qquad (4.36)$$

Tempo de viagem

(i) Atendimentos tipo 1:

As expressões relativas ao tempo de viagem de atendimentos tipo 1 (despacho de um único servidor) no caso de SAEs como o do exemplo 4, são as mesmas apresentadas para o SAE do exemplo 3. Por exemplo:

Tempo médio de viagem no sistema para chamadas tipo 1:

$$\overline{T}^{[1]} = \sum_{j=1}^{N}\sum_{i=1}^{N_A} f_{ji}^{[1]}.t_{ji} \tag{4.37}$$

Tempo médio de viagem para cada servidor $j$, em atendimentos tipo 1:

$$\overline{TU}_j^{[1]} = \frac{\sum_{i=1}^{N_A} f_{ji}^{[1]}.t_{ji}^{[1]}}{\sum_{i=1}^{N_A} f_{ji}^{[1]}} \tag{4.38}$$

Resolvendo estas expressões para o exemplo 4 utilizando os dados de $t_{ji}$ da tabela 4.4 e os resultados de $f_{ji}^{[1]}$ da tabela 4.6, temos que $\overline{T}^{[1]}$ = 5,761 , $\overline{TU}_1^{[1]}$ = 5,997, $\overline{TU}_2^{[1]}$ = 5,710 e $\overline{TU}_3^{[1]}$ = 5,997 minutos.

(ii) Atendimentos tipo 2:

Tempo médio de viagem no sistema para chamadas do tipo 2 (considerando o primeiro veículo a chegar no sistema), como a expressão (3.50):

$$\overline{T}^{[2]} = \sum_{i=1}^{N_A}\left[\left[\sum_{j=1}^{N-1}\sum_{k=j+1}^{N} f_{(j,k)i}^{[2]} \min(t_{ji}, t_{ki})\right] + \sum_{j=1}^{N} f_{ji}^{[2]} t_{ji}\right] \tag{4.39}$$

Tempo médio de viagem no sistema para chamadas tipo 2 (considerando todos os veículos que viajam para atender um chamado do tipo 2), como a expressão (3.52):

$$\overline{T}_t^{[2]} = \sum_{i=1}^{N_A}\left[\left[\sum_{j=1}^{N-1}\sum_{k=j+1}^{N} f_{(j,k)i}^{[2]} (t_{ji} + t_{ki})\right] + \sum_{j=1}^{N} f_{ji}^{[2]} t_{ji}\right] \tag{4.40}$$

No caso de SAE em rodovia do exemplo 4, nas equações (4.39) e (4.40) o valor de $f_{(j,k)i}^{[2]}$ somente é diferente de zero para um único par de servidores $j$ e $k$ da lista de preferência do átomo $i$. Resolvendo a expressão para este exemplo, temos: $\overline{T}^{[2]}$ = 5,761 e $\overline{T}_t^{[2]}$ = 10,374 minutos.

Tempo médio de viagem para o servidor preferencial das chamadas tipo 2, como na expressão (3.52) :

$$\sum_{i=1}^{N_A}\sum_{j=1}^{N}\left[ f_{ji}^{[2]} + \sum_{k\in L'_{ji}} f_{(j,k)i}^{[2]} \right] t_{ji} \qquad (4.41)$$

onde $L'_{ji}$ corresponde ao conjunto de servidores que possuem menor prioridade que o servidor *j* na lista de despacho do átomo $i$. No caso do exemplo 4, $L'_{ji}$ é composto somente pelo servidor *backup* de cada átomo $i$ (p.e, $L'_{22} = \{1\}$ e $L'_{12} = \phi$). O valor da expressão (4.41), para os dados do exemplo 4 é 3,857 minutos.

Tempo médio de viagem para o servidor *backup* das chamadas tipo 2:

Na expressão (3.53) para o exemplo 2, considera-se que o servidor *backup* $k$ é aquele que participa de um atendimento tipo 2 e é despachado como segunda, terceira,..., opção da lista de servidores disponíveis, dado que qualquer servidor pode atender a qualquer átomo. Como no exemplo 4, há uma política particular de despacho, onde apenas alguns servidores podem atender um certo átomo do sistema, propomos a expressão (4.42) a seguir:

$$\sum_{i=1}^{N_A}\sum_{j=1}^{N}\left[ \sum_{k\in L'_{ji}} f_{(j,k)i}^{[2]} + f_{ki}^{[2]} \right] t_{ki} \qquad (4.42)$$

Na expressão (4.42) consideramos os atendimentos de chamadas tipo 2 em que o servidor *backup* $k$ é o único a ser despachado ($f_{ki}^{[2]}$) e os atendimentos realizados pelo servidor *backup* $k$ com o servidor preferencial *j* do átomo *i* ($f_{(j,k)i}^{[2]}$).

Diferentemente dos sistemas de patrulha policial, os servidores dos SAEs em rodovias estão fixos em suas bases quando disponíveis. Desta forma, na maioria dos casos, o primeiro servidor a chegar no local do acidente corresponde ao primeiro servidor preferencial (servidor mais próximo). No entanto, o inverso também pode acontecer, dependendo das extensões das áreas primárias de cada átomo, das condições de tráfego nas estradas e outros fatores.

Tempo médio de viagem para o primeiro servidor a chegar no local de uma chamada tipo 2:

$$\overline{T}_F = \frac{\sum_{i=1}^{N_A}\sum_{j=1}^{N-1}\sum_{k=j+1}^{N} f_{(j,k)i}^{[2]} \min(t_{ji}, t_{ki})}{\sum_{j=1}^{N-1}\sum_{k=j+1}^{N} f_{(j,k)i}^{[2]}} \tag{4.43}$$

Tempo médio de viagem para o segundo servidor a chegar no local de uma chamada tipo 2:

$$\overline{T}_S = \frac{\sum_{i=1}^{N_A}\sum_{j=1}^{N-1}\sum_{k=j+1}^{N} f_{(j,k)i}^{[2]} \max(t_{ji}, t_{ki})}{\sum_{j=1}^{N-1}\sum_{k=j+1}^{N} f_{(j,k)i}^{[2]}} \tag{4.44}$$

Estas equações somente consideram os atendimentos em que participam dois servidores (ambulâncias). Note que, as equações (4.43) e (4.44) diferem das equações (3.54) e (3.56), pois nos SAEs em rodovias os servidores estão fixos em suas bases, e não é necessário considerar a probabilidade do servidor estar em determinado átomo do sistema. Os valores de $\overline{T}_F$ e $\overline{T}_S$ no exemplo 4 são 5,0 e 8,5 minutos, respectivamente. Propomos também a expressão para calcular o tempo médio de viagem para cada servidor $j$, em atendimentos tipo 2:

$$\overline{TU}_j^{[2]} = \frac{\sum_{i=1}^{N_A}\left[\sum_{k\neq j}^{N} f_{(j,k)i}^{[2]} t_{ji} + f_{ji}^{[2]} t_{ji}\right]}{\sum_{i=1}^{N_A}\left[\sum_{k\neq j}^{N} f_{(j,k)i}^{[2]} + f_{ji}^{[2]}\right]} \tag{4.45}$$

Resolvendo a expressão (4.46) para os 3 servidores do exemplo 4, obtemos : $\overline{TU}_1^{[2]}$ = 5,997, $\overline{TU}_2^{[2]}$ = 5,710 e $\overline{TU}_3^{[2]}$ = 5,997 minutos.

Fração de chamadas atendidas em mais que $\bar{t}_v$ minutos:

Podemos também calcular a probabilidade de uma chamada no sistema ser atendida em um tempo superior a $\bar{t}_v$ (limite predeterminado), utilizando a distribuição do tempo de viagem da ambulância de sua base ao átomo (local da chamada):

$$P_{t>\bar{tv}} = \sum_{j=1}^{N}\sum_{i=1}^{N_A} f_{ji}^{[v]}\,, \tag{4.46}$$

onde $f_{ij}^{[v]}$ corresponde a fração de todas as chamadas para as quais o servidor *j* é enviado ao átomo *i*, e cujo tempo de viagem excede $\bar{t}_v$. Esta medida pode ser definida como:

$$f_{ji}^{[v]} = p\left(t_{ji} > \bar{t}_v\right)\left(f_{ji}^{'[1]} + f_{ji}^{'[2]} + \sum_{k\in L_i^{[v]}} f_{(j,k)i}^{'[2]}\right), \tag{4.47}$$

onde $L_i^{[v]}$ corresponde ao conjunto de servidores da lista de despacho do átomo *i* tal que $t_{ji} < t_{ki}$, e o termo $p\left(t_{ji} > \bar{t}_v\right)$ corresponde a fração de chamadas no átomo *i* que esperam mais que $\bar{t}_v$ min pela chegada do servidor *j*.

No presente estudo, este modelo foi aplicado ao SAE *Centrovias* e os resultados são apresentados no capítulo 6.

## 4.3 Adaptação do modelo Hipercubo múltiplo despacho para análise dos SAEs em rodovias, considerando um terceiro estado do servidor:

O modelo proposto por LARSON & MCKNEW (1982) comentado na seção 3.2.5 do capítulo 3 pode ser adaptado para análise dos SAEs em rodovias considerando que nestes sistemas, parte das chamadas atendidas pelos servidores são atendidas na sua base e possuem características diferenciadas com relação às chamadas que decorrem de acidentes ao longo da rodovia. Por exemplo, um usuário da rodovia ou trabalhador de uma instalação próxima da base do SAE na rodovia pode solicitar atendimento na própria base. Tais eventos tornam os servidores ocupados e são caracterizados por tempo de viagem igual a zero, conseqüentemente o tempo médio de atendimento deve ser diferenciado para estas chamadas.

Outra característica importante é que se o servidor estiver ocupado quando este tipo de chamada chega na sua base, a chamada não é atendida pelo servidor *backup*, dado que não se trata de um atendimento de emergência e a chamada é perdida para o sistema. Como descrito anteriormente no capítulo 3, LARSON & MCKNEW (1982) propuseram uma extensão do modelo Hipercubo, para estudar os sistemas de patrulhamento policial considerando que, uma parte do tempo que permanecem ocupadas, as viaturas de policia estão atendendo eventos que não foram atribuídos por ordem de despacho da central. Os autores chamam estes eventos de PIAs (*Patrol Initiated Activities*). Nesta seção apresentamos como aquele modelo pode ser adaptado para estudar os SAEs em rodovias considerando chamadas que são diferenciadas por serem atendidas na base do servidor. Neste estudo denominamos estas chamadas como tipo 1a (atendimento na base). Outras características dos SAEs em rodovias devem ser mantidas, tais como:

- Duplo despacho de servidores para atender um mesmo acidente;
- Não há fila de espera e a prob1ailidade de perda é uma medida importante, sendo que uma chamada pode ser perdida mesmo se há servidores disponíveis;
- A política de despacho deve considerar que somente alguns servidores do sistema podem atender um determinado átomo (*backup* parcial).

Apresentamos a seguir como o sistema em rodovia do exemplo 3 e 4, pode ser analisado considerando múltiplo despacho e chamadas do tipo 1a. Para chamadas do tipo 1a temos: $\lambda_i^{[1a]}$ (taxa de chegada de chamadas tipo 1a no átomo *i*).

**Exemplo ilustrativo 5**: sistema SAE em rodovia c/ múltiplo despacho e chamadas 1a

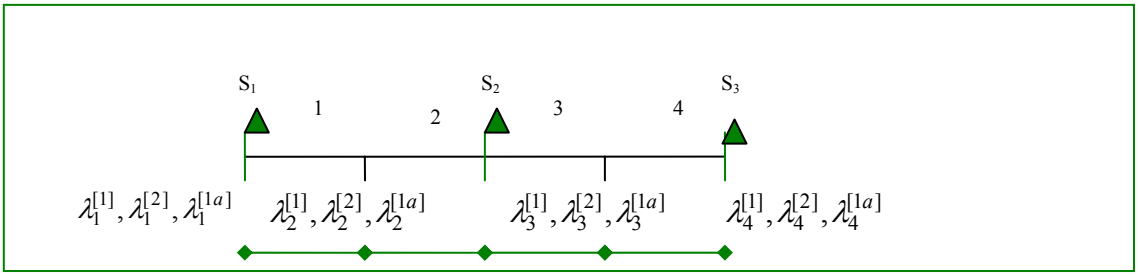


Figura 4.4 : SAE em rodovia - múltiplo despacho c/ chamadas tipo 1a - Exemplo 5 (com $N_A$= 4 átomos e N = 3 servidores)

O sistema da figura 4.4 possui as mesmas características do sistema do exemplo 4, com relação ao número e disposição dos servidores (N = 3 servidores), distribuição dos átomos ($N_A$ = 4 átomos), lista de preferência de despacho para cada átomo (tabela 4.1) e chamadas do tipo 1 (único despacho) e tipo 2 (múltiplo despacho). No entanto, neste sistema, temos que, em cada átomo $i$, pode ocorrer chamadas tipo 1a ($\lambda_i^{[1a]}$), que são diferenciadas e ocorrem na própria base dos servidores. E cada servidor $j$ possui a taxa de atendimento $\mu_j^{[I]}$ para chamadas tipo 1 e 2 e $\mu_j^{[II]}$ para chamadas do tipo 1a em sua base.

Com relação à política de despacho, uma chamada tipo 1 pode ser atendida por um servidor preferencial ou um *backup*. Uma chamada tipo 2 é atendida por dois servidores ao mesmo tempo, e é perdida pelo sistema se os dois servidores preferenciais estiverem ocupados. Se uma chamada tipo 2 ocorre quando um dos dois servidores estiver ocupado, o outro disponível deve atender sem ajuda de um terceiro servidor do sistema. No caso de chamadas do tipo 1a, somente o servidor preferencial atende a chamada que chega na sua base. Se o mesmo estiver ocupado, a chamada é perdida para o sistema.

**Transição de estados:**

Como discutimos anteriormente, ao consideramos chamadas do tipo 1a no sistema, cada servidor passa a ter 3 estados possíveis, dado que estas chamadas requerem atendimento diferenciado das chamadas de emergência recebidas pela central. Assim, um servidor pode estar nos estados: (0) livre; (1) ocupado atendendo uma chamada de emergência ao longo da rodovia; (2) ocupado atendendo uma chamada do tipo 1a na sua base. Desta forma, há $3^3 = 27$ estados possíveis do sistema, que são:

$\{000\},\{001\},\{010\},\{011\},\{100\},\{101\},\{110\},\{111\},\{002\},\{020\},$
$\{022\},\{200\},\{202\},\{220\},\{222\},\{112\},\{121\},\{122\},\{211\},\{212\},\{221\},\{120\},\{102\},$
$\{201\},\{210\},\{012\},\{021\}.$

Note que, neste exemplo novas transições devem ser consideradas ao analisarmos um sistema com chamadas do tipo 1a pelo modelo Hipercubo. Exemplos de novas transições são: $\{110\} \rightarrow \{112\}$; $\{000\} \rightarrow \{020\}$; $\{020\} \rightarrow \{000\}$; $\{210\} \rightarrow \{212\}$; $\{002\} \rightarrow \{112\}$ e outras transições, sendo que o servidor pode passar de livre (0) para outros dois estados (1) e (2) e destes estados para livre.

As transições em que o servidor sai do estado livre (0) para o estado ocupado atendendo uma chamada do tipo 1a, ocorrem com a chegada de uma chamada do tipo 1a nos átomos atendidos preferencialmente por este servidor. No exemplo 5, por exemplo, chamadas do tipo 1a que chegam nos átomos 2 ou 4 são atendidas pelo servidor 2 somente se o mesmo estiver livre (veja tabela 4.1), provocando as transições:

$\{000\} \rightarrow \{020\}$ ; $\{100\} \rightarrow \{120\}$ ; $\{200\} \rightarrow \{220\}$ ; $\{001\} \rightarrow \{021\}$ ; $\{002\} \rightarrow \{022\}$ ;
$\{101\} \rightarrow \{121\}$ ; $\{201\} \rightarrow \{221\}$ ; $\{102\} \rightarrow \{122\}$ ; $\{202\} \rightarrow \{222\}$ .

**Equações de equilíbrio:**

A seguir, apresentamos uma breve discussão das equações de transição entre os estados do sistema. Admitimos que $\lambda^{[1]} = \sum_{i=1}^{N_A} \lambda_i^{[1]}$ , $\lambda^{[2]} = \sum_{i=1}^{N_A} \lambda_i^{[2]}$ , $\lambda^{[1a]} = \sum_{i=1}^{N_A} \lambda_i^{[1a]}$ $\lambda = \lambda^{[1]} + \lambda^{[2]} + \lambda^{[1a]}$ e $\mu^{[I]} = \sum_{j=1}^{N} \mu_j^{[I]}$ e $\mu^{[II]} = \sum_{j=1}^{N} \mu_j^{[II]}$

(v) Quando todos os servidores estão disponíveis: {000}

Neste caso a equação é similar a equação do modelo Hipercubo com múltiplo despacho, mas consideramos também chamadas do tipo 1a, sendo $\lambda = \lambda^{[1]} + \lambda^{[2]} + \lambda^{[1a]}$ temos:

$$(\lambda).P_{\{000\}} = \mu_1^{[1]}.P_{\{100\}} + \mu_2^{[I]}.P_{\{010\}} + \mu_3^{[I]}.P_{\{001\}} + \mu_1^{[II]}.P_{\{200\}} + \mu_2^{[II]}.P_{\{020\}} + \mu_3^{[II]}.P_{\{002\}} \quad (4.48)$$

(vi)Quando há apenas 1 servidor ocupado atendendo chamada de emergência no sistema:

As equações de equilíbrio dos estados {001}, {010} e {100} no exemplo 5 são similares às equações de equilíbrio destes estados nos demais exemplos, mas outras transições devem ser adicionadas, por exemplo o estado {010}:

{010}
$$(\lambda - (\lambda_2^{[1a]} + \lambda_3^{[1a]}) + \mu_2^{[I]}).P_{\{010\}} = (\lambda_2^{[1]} + \lambda_3^{[1]}).P_{\{000\}} + \mu_1^{[I]}.P_{\{110\}} + \mu_3^{[I]}.P_{\{011\}} + \mu_1^{[II]}.P_{\{210\}} + \mu_3^{[II]}.P_{\{012\}}$$
(4.49)

Note que, chamadas do tipo 1a requerem um tempo médio de atendimento diferenciado, o que faz com que outras transições para dentro e fora do estado {010} sejam consideradas. Neste

exemplo, chamadas do tipo 1a nos átomos 1 e 4 são atendidas pelos seus servidores preferenciais (1 e 3, respectivamente). Por outro lado, chamadas do tipo 1a que chegam nos átomos 2 e 3 não são atendidas pois o seu servidor preferencial 2 está ocupado atendendo uma chamada de emergência.(chamada do tipo 1 ou tipo 2)

(vii) Quando há 2 servidores ocupados no sistema atendendo chamadas do tipo 1 ou 2 (chamadas de emergência)

Analisando por exemplo o estado $\{110\}$, temos a equação de equilíbrio:

$$(\lambda_3^{[1]}+\lambda_4^{[1]}+\lambda_3^{[2]}+\lambda_4^{[2]}+\lambda_4^{[1a]}+\mu_1^{[I]}+\mu_2^{[I]}).P_{\{110\}}=(\lambda_1^{[2]}+\lambda_2^{[2]}).P_{000}+(\lambda_1^{[1]}+\lambda_2^{[1]}+\lambda_1^{[2]}+\lambda_2^{[2]}).P_{\{010\}}+$$
$$(\lambda_1^{[1]}+\lambda_2^{[1]}+\lambda_3^{[1]}+\lambda_1^{[2]}+\lambda_2^{[2]}).P_{\{100\}}+\mu_3^{[I]}.P_{\{111\}}+\mu_3^{[II]}.P_{\{112\}} \quad (4.50)$$

Lado esquerdo da equação (4.50) – fluxo para fora do estado $\{110\}$:

7. $\{110\} \rightarrow \{111\}$ – ocorre com a chegada de uma chamada do tipo 1 ou tipo 2, nos átomos 3 e 4 (taxa total $\lambda_3^{[1]}+\lambda_4^{[1]}+\lambda_3^{[2]}+\lambda_4^{[2]}$).
8. $\{110\} \rightarrow \{112\}$ – ocorre com a chegada de uma chamada do tipo 3 no átomo 4 (taxa $\lambda_4^{[1a]}$), pois somente o servidor 3 está livre e pode atender chamadas neste átomo, sendo o servidor preferencial;
9. $\{110\} \rightarrow \{010\}$ – ocorre com o término de serviço do servidor 1 (taxa $\mu_1^{[I]}$);
10. $\{110\} \rightarrow \{100\}$ – ocorre com o término de serviço do servidor 2 (taxa $\mu_2^{[II]}$).

Lado direito da equação (4.50) – fluxo para dentro do estado $\{110\}$:

9. $\{111\} \rightarrow \{110\}$ – ocorre com o término de serviço do servidor 3 (taxa $\mu_3^{[I]}$);
10. $\{112\} \rightarrow \{110\}$ – ocorre com o término de serviço do servidor 3 em uma chamada do tipo 3 no átomo 4(taxa $\mu_3^{[II]}$);
11. $\{010\} \rightarrow \{110\}$ – ocorre com a chegada de uma chamada tipo 1 ou tipo 2 no átomo 1 (taxa $\lambda_1^{[1]}+\lambda_1^{[2]}$) ou no átomo 2 (taxa $\lambda_2^{[1]}+\lambda_2^{[2]}$);

12. {100} → {110} – ocorre com a chegada de uma chamada do tipo 1 nos átomos 1, 2 ou 3 (taxa total $\lambda_1^{[1]} + \lambda_2^{[1]} + \lambda_3^{[1]}$) e do tipo 2 nos átomos 1 e 2 (taxa total $\lambda_1^{[2]} + \lambda_2^{[2]}$), como na equação (4.17);
13. {000} → {110} – ocorre com a chegada de uma chamada do tipo 2 nos átomos 1 ou 2 (taxa total $\lambda_1^{[2]} + \lambda_2^{[2]}$ ), como no exemplo 2.

(viii) Quando todos os servidores estão ocupados atendendo chamadas do tipo 1 ou 2 (chamadas de emergência): {111}

$$\mu^{[I]}.P_{\{111\}} = (\lambda_1^{[1]} + \lambda_2^{[1]} + \lambda_1^{[2]} + \lambda_2^{[2]}).P_{\{011\}} + \lambda.P_{\{101\}} + (\lambda_3^{[1]} + \lambda_4^{[1]} + \lambda_3^{[2]} + \lambda_4^{[2]}).P_{\{110\}} + (\lambda_1^{[2]} + \lambda_2^{[2]}).P_{\{001\}} + (\lambda_3^{[2]} + \lambda_4^{[2]}).P_{\{100\}} \quad (4.51)$$

Observe que, a equação (4.51) não difere da equação (4.19) do exemplo 4.

(ix)Quando há apenas 1 servidor ocupado atendendo chamadas do tipo 1A:

Estes estados são {002}, {020} e {200}, tomando como exemplo o estado {020}, temos:

$$\{020\}\ (\lambda - (\lambda_2^{[1a]} + \lambda_3^{[1a]}) + \mu_2^{[II]}).P_{\{020\}} = (\lambda_2^{[1a]} + \lambda_3^{[1a]}).P_{\{000\}} + \mu_1^{[I]}.P_{\{120\}} + \mu_3^{[I]}.P_{\{021\}} + \mu_1^{[II]}.P_{\{220\}} + \mu_3^{[II]}.P_{\{022\}} \quad (4.52)$$

Note no lado esquerdo da equação (4.52) que chamadas do tipo 1a nos átomos 2 e 3 não podem ser atendidas, pois seu servidor preferencial (2) está ocupado atendendo uma chamada do tipo 1a. Chamadas do tipo 1 e 2 podem ser atendidas pelos servidores *backup*, conforme discutido anteriormente.

No lado direito da equação, temos as seguintes transições para dentro do estado {020}:

1. {120} → {020} – ocorre com o término de serviço do servidor 1 ao atender uma chamada do tipo 1 ou tipo 2 (taxa $\mu_1^{[I]}$);
2. {021} → {020} – ocorre com o término de serviço do servidor 3 ao atender uma chamada do tipo 1 ou tipo 2 (taxa $\mu_3^{[I]}$);

3. {220} → {020} – ocorre com o término de serviço do servidor 1 ao atender uma chamada do tipo 3 no átomo 1 (taxa $\mu_1^{[II]}$);
4. {022} → {020} – ocorre com o término de serviço do servidor 3 ao atender uma chamada do tipo 1a no átomo 4 (taxa $\mu_3^{[II]}$);
5. {000} → {020} – ocorre com a chegada de uma chamada do tipo 1a nos átomos 2 ou 3 (taxa total $\lambda_2^{[1a]} + \lambda_3^{[1a]}$);

(x) Quando há 2 servidores ocupados no sistema atendendo chamadas do tipo 1a:
Por exemplo para o estado {220}, temos a equação de equilíbrio:

$$(\lambda_3^{[1]} + \lambda_4^{[1]} + \lambda_3^{[2]} + \lambda_4^{[2]} + \lambda_4^{[1a]} + \mu_1^{[II]} + \mu_2^{[II]}).P_{\{220\}} = (\lambda_1^{[1a]}).P_{\{020\}} + (\lambda_2^{[1a]} + \lambda_3^{[1a]}).P_{\{200\}} + \mu_3^{[I]}.P_{\{221\}} + \mu_3^{[II]}P_{\{222\}} \quad (4.53)$$

Analisando as possíveis transições para fora deste estado na equação (4.53), temos:

1. {220} → {221} – ocorre com a chegada de uma chamada do tipo 1 ou tipo 2, nos átomos 3 e 4 (taxa total $\lambda_3^{[1]} + \lambda_4^{[1]} + \lambda_3^{[2]} + \lambda_4^{[2]}$).
2. {220} → {222} – ocorre com a chegada de uma chamada do tipo 1a no átomo 4 (taxa $\lambda_4^{[1a]}$);
3. {220} → {020} – ocorre com o término de serviço do servidor 1 ao atender uma chamada do tipo 1a (taxa $\mu_1^{[II]}$);
4. {220} → {200} – ocorre com o término de serviço do servidor 2 ao atender uma chamada do tipo 1a (taxa $\mu_2^{[II]}$);

Lado direito da equação (4.53) – fluxo para dentro do estado {220}:

1. {221} → {220} – ocorre com o término de serviço do servidor 3 (taxa $\mu_3^{[I]}$);
2. {222} → {220} – ocorre com o término de serviço do servidor 3 em uma chamada do tipo 1a no átomo 4 (taxa $\mu_3^{[II]}$);
3. {020} → {220} – ocorre com a chegada de uma chamada tipo 1a no átomo 1 (taxa $\lambda_1^{[1a]}$) ;

4. $\{200\} \rightarrow \{220\}$ – ocorre com a chegada de uma chamada do tipo 1a nos átomos 2 ou 3 (taxa total $\lambda_2^{[1a]} + \lambda_3^{[1a]}$) , dado que o servidor 2 é o servidor preferencial;

(xi)Quando há 2 servidores ocupados no sistema: sendo um servidor atendendo chamadas do tipo 1 ou 2 (chamadas de emergência) e um servidor ocupado atendendo chamadas do tipo 1a:

Analisando por exemplo o estado $\{210\}$, temos a equação de equilíbrio:

$$(\lambda_3^{[1]} + \lambda_4^{[1]} + \lambda_3^{[2]} + \lambda_4^{[2]} + \lambda_4^{[1a]} + \mu_1^{[II]} + \mu_2^{[I]}).P_{\{210\}} = (\lambda_1^{[1a]}).P_{\{010\}} + (\lambda_1^{[1]} + \lambda_2^{[1]} + \lambda_3^{[1]} + \lambda_1^{[2]} + \lambda_2^{[2]}).P_{\{200\}} + \mu_3^{[I]}.P_{\{211\}} + \mu_3^{[II]}.P_{\{212\}} \quad (4.54)$$

Lado esquerdo da equação (4.54) – fluxo para fora do estado $\{210\}$:

1. $\{210\} \rightarrow \{211\}$ – ocorre com a chegada de uma chamada do tipo 1 ou tipo 2, nos átomos 3 e 4 (taxa total $\lambda_3^{[1]} + \lambda_4^{[1]} + \lambda_3^{[2]} + \lambda_4^{[2]}$).
2. $\{210\} \rightarrow \{212\}$ – ocorre com a chegada de uma chamada do tipo 1a no átomo 4 (taxa $\lambda_4^{[1a]}$), atendida pelo servidor 3;
3. $\{210\} \rightarrow \{010\}$ – ocorre com o término de serviço do servidor 1 ao atender uma chamada do tipo 1a (taxa $\mu_1^{[II]}$);
4. $\{210\} \rightarrow \{200\}$ – ocorre com o término de serviço do servidor 2 ao atender uma chamada do tipo 1 ou tipo 2 – chamadas de emergência (taxa $\mu_2^{[I]}$).

Lado direito da equação (4.54) – fluxo para dentro do estado $\{210\}$:

1. $\{211\} \rightarrow \{210\}$ – ocorre com o término de serviço do servidor 3 ao atender uma chamada do tipo 1 ou tipo 2 (taxa $\mu_3^{[I]}$);
2. $\{212\} \rightarrow \{210\}$ – ocorre com o término de serviço do servidor 3 em uma chamada do tipo 1a no átomo 4(taxa $\mu_3^{[II]}$);
3. $\{010\} \rightarrow \{210\}$ – ocorre com a chegada de uma chamada tipo 1a no átomo 1 (taxa $\lambda_1^{[2]}$);

4. $\{200\} \rightarrow \{210\}$ – ocorre com a chegada de uma chamada do tipo 1 nos átomos 1, 2 ou 3 (taxa total $\lambda_1^{[1]} + \lambda_2^{[1]} + \lambda_3^{[1]}$) e do tipo 2 nos átomos 1 e 2 (taxa total $\lambda_1^{[2]} + \lambda_2^{[2]}$);

(xii) Quando todos os servidores estão ocupados atendendo chamadas do tipo 3: {222}

$$\mu.^{[II]} P_{\{222\}} = (\lambda_1^{[1a]}).P_{\{022\}} + (\lambda_2^{[1a]} + \lambda_3^{[1a]}).P_{\{202\}} + (\lambda_4^{[1a]}).P_{\{220\}} \qquad (4.55)$$

Note que, as transições para fora do estado {222} requerem que um dos servidores ocupados termine de atender uma chamada do tipo 1a. O lado direito da equação (4.55) apresenta as transições em cada servidor do sistema pode torna-se ocupado ao chegar uma chamada do tipo 1a em sua base (i.e, estados {022}, {202} e {220}).

**Medidas de Desempenho:**

Carga de trabalho dos servidores

Ao considerar o atendimento de chamadas que não são de emergência (chamadas tipo 1a), podemos calcular pelo modelo Hipercubo a carga de trabalho de cada servidor atendendo cada tipo de chamada. Assim temos: $\rho_j^{[I]}$ é a carga de trabalho do servidor *j* atendendo uma chamada de emergência (tipo 1 ou tipo 2) e $\rho_j^{[II]}$ é a carga de trabalho do servidor *j* atendendo uma chamada tipo 1a. Para o exemplo 5 com três servidores temos:

Chamadas tipo 1 e 2:

$$\rho_1^{[I]} = P_{100} + P_{101} + P_{102} + P_{110} + P_{111} + P_{112} + P_{120} + P_{121} + P_{122}$$
$$\rho_2^{[I]} = P_{010} + P_{011} + P_{012} + P_{110} + P_{111} + P_{112} + P_{210} + P_{211} + P_{212}$$
$$\rho_3^{[I]} = P_{001} + P_{011} + P_{101} + P_{021} + P_{111} + P_{121} + P_{201} + P_{211} + P_{221}$$

Chamadas tipo 1a:

$$\rho_1^{[II]} = P_{200} + P_{201} + P_{202} + P_{210} + P_{211} + P_{212} + P_{220} + P_{221} + P_{222}$$
$$\rho_2^{[II]} = P_{020} + P_{021} + P_{022} + P_{120} + P_{121} + P_{122} + P_{220} + P_{221} + P_{222}$$
$$\rho_3^{[II]} = P_{002} + P_{012} + P_{022} + P_{102} + P_{112} + P_{122} + P_{202} + P_{212} + P_{222}$$

Probabilidade de Perda

Como discutido anteriormente, a probabilidade de perda é uma medida importante para os SAEs, pois uma chamada pode ser perdida mesma quando há servidores disponíveis. No caso de SAEs em rodovias com política de múltiplo despacho e ocorrência de chamadas 1a, podemos calcular: probabilidade de perda de chamadas tipo 1 ($P_p^{[1]}$), probabilidade de perda de chamadas tipo 2 ($P_p^{[2]}$), probabilidade de perda de chamadas tipo 1a ($P_p^{[1a]}$) e probabilidade de perda para qualquer chamada do sistema ($P_p$), de forma similar a discutida na seção 4.2. No caso de chamadas do tipo 1a, a perda ocorre quando o servidor preferencial estiver ocupado, pois não há atendimento *backup*.

Freqüências de despacho

As medidas de freqüência de despacho de cada servidor a cada átomo para chamadas tipo 1 e tipo 2 são similares as equações 4.25 a 4.36. No presente modelo, a expressão 3.47 que calcula freqüência total de chamadas no sistema atendidas pelo servidor *j* no átomo *i*, deve redefinida para considerar também chamadas tipo 1a.

Freqüência de despachos do servidor $j$ ao átomo $i$ (considerando todos os despachos e os três tipos de chamadas):

$$f_{ji} = \frac{\frac{\lambda_i^{[1]}}{\lambda} \sum_{B \in E_{ji}} P_B + \frac{\lambda_i^{[2]}}{\lambda} \sum_{B \in E_{(j,k)i}} P_B + \frac{\lambda_i^{[2]}}{\lambda} P_{B \in F_{ji}} + \frac{\lambda_i^{[1a]}}{\lambda} \sum_{B \in G_{ji}} P_B}{(1 - P_p)} \qquad (4.56)$$

Na expressão (4.56) acima, $E_{ji}, E_{(j,k)i}$ e $F_{ji}$ foram definidos nas seções anteriores do capítulo 3 e 4, e $G_{ji}$ corresponde ao conjunto de estados nos quais o servidor preferencial *j* do átomo *i* está disponível. Por exemplo, no exemplo 5, temos:
$G_{23} = \{\{000\},\{100\},\{001\},\{101\},\{200\},\{002\},\{202\},\{102\},\{201\}\}$. Note que, como na expressão (4.36), temos: $\sum_{j=1}^{N} \sum_{i=1}^{N_A} f_{ji} = 1$.

As medidas de tempo de viagem são similares as expressões 4.37 a 4.45, descritas na seção anterior para o exemplo 4.

### 4.4 Modelo Hipercubo múltiplo despacho aplicado a sistemas médico emergencial em rodovias com servidores diferenciados (carro médico):

Alguns SAEs em rodovias operam com servidores diferenciados (além de chamadas diferenciadas do tipo 1 e 2), sendo que o despacho do tipo de servidor depende do tipo de chamado. Por exemplo, alguns destes SAEs possuem veículos resgates, que são veículos em geral mais pesados que transportam, além de equipamentos de primeiros socorros, equipamentos para combate a incêndio e quebra de ferragens para remoção das vítimas de acidente. Outro tipo de veículo utilizado é a viatura UTI, que transporta equipamentos médicos mais especializados para tratamento das vítimas durante o transporte para o hospital, porém em geral não transporta equipamentos de combate a incêndio. Além disso, outra possível diferença entre estes dois tipos de veículos é a equipe de profissionais transportada (resgatistas, médicos, enfermeiros). Esta situação é semelhante a dos sistemas urbanos SAMU, descrita em TAKEDA et al. (2000, 2004), onde as chamadas podem ser do tipo avançada (muito urgentes) ou básicas (urgentes) e dependendo do tipo de chamada é despachado prioritariamente um veículo de atendimento avançado (VSA), ou um veículo de atendimento básico (VSB).

Há SAEs que possuem também um veículo que transporta apenas o médico ao local do acidente (carro médico) e é despachado simultaneamente com o veículo resgate. O carro médico não transporta pacientes, medicamentos ou equipamentos. Em algumas das mais congestionadas rodovias do Brasil, os SAEs utilizam também helicópteros para socorro das vítimas para determinados tipos de acidente.

Além das demais características de SAEs com múltiplo despacho, *backup* parcial e servidores idênticos (exemplo 4), devemos também considerar que nestes sistemas:

- As chamadas em cada átomo são diferenciadas (p.e., tipo 1,2,3..) e requerem diferentes tipos de servidores, portanto cada tipo de chamada em cada átomo possui uma lista de preferência de despacho. Assim, um dado átomo possui mais de uma lista de preferência. Note que esta situação é diferente das anteriores, todas com mesma lista de despacho por átomo.

- A política de despacho torna-se mais complexa dado que podemos ter chamadas que requerem: 1 servidor, 2 servidores idênticos, 2 servidores diferenciados, 3 servidores idênticos, 3 servidores diferenciados, etc.

- Além disso, a política de *backup* parcial também torna-se mais complexa dado que podemos ter chamadas com até 2, 3,..., servidores candidatos em sua lista de despacho.

A seguir apresentamos um exemplo de um SAE em rodovias que é similar ao estudo de caso desta pesquisa discutido no capítulo 2 (*Centrovias* 2) e, cuja análise computacional é apresentada no capítulo 6. Neste sistema há dois tipos de veículos: carro resgate e carro médico. Cada átomo do sistema possui chamadas de 4 tipos:

- chamadas tipo 1 que requerem um único despacho (um veículo resgate), dispondo para isso de até 2 possíveis servidores (como no exemplo 4).
- chamadas tipo 2a que requerem duplo despacho (dois veículos resgates idênticos), dispondo de até 2 possíveis servidores (como no exemplo 4);
- chamadas tipo 2b que requerem duplo despacho (dois veículos diferenciados: 1 veículo resgate e o carro médico), dispondo de até 3 possíveis servidores (há um terceiro veículo resgate que é *backup* de um dos dois primeiros, caso esteja ocupado);
- chamadas tipo 3 que requerem triplo despacho (dois veículos resgate e o carro médico, ou três veículos resgates), dispondo de até 3 possíveis servidores (não há *backup* para este tipo de chamada);

Note que, consideramos neste sistema que o carro médico pode ter como *backup* um carro resgate (chamadas tipo 2b), o que ocorre de fato no sistema real analisado. Esta hipótese também torna mais simples a análise. Caso contrário, teríamos que considerar que cada tipo de servidor (carro médico e resgate) teria uma lista de servidores *backup* em cada átomo para cada tipo de chamada.

**Exemplo ilustrativo 6**: sistema SAE em rodovia c/ múltiplo despacho e servidores diferenciados (carro médico e veículos resgate)

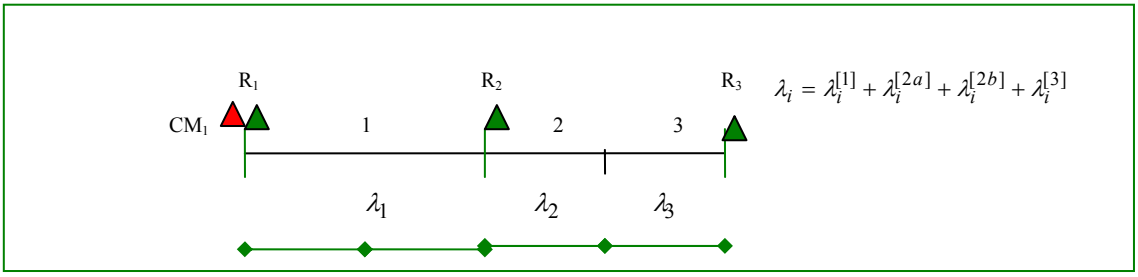


Figura 4.5 : SAE em rodovia - múltiplo despacho c/ carro médico - Exemplo 6 (com $N_A$= 3 átomos e N = 4 servidores).

Por exemplo, considere o sistema exemplo da figura 4.5, com 3 átomos e 4 servidores: 3 resgates (R1,R2 e R3) e 1 carro médico (CM). O carro médico divide sua base com o resgate 1 no átomo 1, o resgate 2 está localizado no átomo 2 e o resgate 3 no átomo 3. Em cada átomo *i,* chamadas podem ser do tipo 1, 2a, 2b e 3, com taxas $\lambda_i^{[1]}$, $\lambda_i^{[2a]}$, $\lambda_i^{[2b]}$ e $\lambda_i^{[3]}$. Para representar os diferentes tipos de chamadas que possuem diferentes listas de preferência de despacho, subdividimos cada átomo de acordo com o número de listas por átomo. No caso deste exemplo, cada átomo é subdividido em duas camadas (*"layering"*): camada *a* para chamadas que são atendidas somente por veículos resgate e camada *b* para chamadas que são atendidas por carro médico e veículos resgate. A lista de preferência de despacho de cada sub-átomo (camada do átomo) é dada a seguir na tabela 4.8, considerando servidor 1 (carro médico), servidor 2 (resgate 1), servidor 3 (resgate 2) e servidor 4 (resgate 3). Note que chamadas do tipo 3 ocorrem nos átomos 1b, 2b e 3a (ao invés de 3b).

Tabela 4.8 – Subdivisão de átomos de acordo com o tipo de chamada e lista de despacho

| Sub-atomo | chamadas | primeiro | segundo | terceiro |
|---|---|---|---|---|
| 1a | 1, 2a | 2 | 3 | - |
| 1b | 2b,3 | 1 | 2 | 3 |
| 2a | 1,2a | 3 | 2 | |
| 2b | 2b,3 | 1 | 3 | 2 |
| 3a | 1,2a,3 | 4 | 3 | 2 |
| 3b | 2b | 1 | 4 | 3 |

Na tabela 4.8, notamos que:

- chamadas do tipo 1 são atendidas pelo servidor preferencial, ou o servidor *backup* se o primeiro estiver ocupado. Se os dois servidores estiverem ocupados, a chamada é perdida para o sistema (átomos 1a, 2a e 3a da tabela 4.8);

- chamadas do tipo 2a são atendidas pelos dois primeiros servidores mais próximos (primeiro e segundo na lista de preferência), despachados simultaneamente. Como descrito anteriormente no exemplo 4, se um dos servidores estiver ocupado, a chamada á atendida pelo outro disponível e se os dois estiverem ocupados, a chamada é perdida para o sistema (átomos 1a, 2a e 3a);
- chamadas do tipo 2b são atendidas pelo primeiro servidor preferencial (sempre o carro médico) e o segundo servidor preferencial (veículo resgate), porém quando um destes estiver ocupado, o terceiro servidor da lista de preferência é despachado. Quando apenas um dos três servidores da lista de preferência de despacho está livre o mesmo é despachado sozinho, podendo receber auxilio de servidores de outro SAE mais próximo. Se os três servidores estiverem ocupados, a chamada é perdida para o sistema (átomos 1b, 2b e 3b);
- chamadas do tipo 3 são atendidas pelo carro médico e dois veículos resgates (átomos 1b e 2b), ou por três veículos resgates (átomo 3a). Se apenas dois dos três servidores estiverem livres os mesmos devem atender a chamada como duplo despacho e se há apenas um disponível o mesmo deve também ser despachado sozinho como descrito acima. Se os três servidores estiverem ocupados, a chamada é perdida para o sistema.

**Transição de estados:**

Como há 4 servidores no sistema sem fila de espera, os $2^4 = 16$ possíveis estados do sistema são:
{0000},{0001},{0010},{0011},{0100},{0101},{0110},{0111},{1000},{1001},{1010},
{1011}, {1100},{1101},{1110}, {1111}.

Entre as transições adicionais com relação ao exemplo 4, estão por exemplo, as transições em que 3 servidores do sistema passam de livres a ocupados. No exemplo 6, estas transições são:
$\{0000\}\rightarrow\{1110\},\{0001\}\rightarrow\{1111\},\{0000\}\rightarrow\{0111\},\{1000\}\rightarrow\{1111\}$.
Outro exemplo são as transições em que o terceiro servidor de um dado átomo atende chamadas do tipo 3, quando os dois primeiros servidores estão ocupados, tais como:,
$\{1011\}\rightarrow\{1111\},\{1101\}\rightarrow\{1111\},\{1010\}\rightarrow\{1011\},\{1100\}\rightarrow\{1110\},\{1101\}\rightarrow\{1111\}$,
$\{1110\}\rightarrow\{1111\},\{1001\}\rightarrow\{1011\},\{0011\}\rightarrow\{1011\},\{0110\}\rightarrow\{1110\}$, $\{0111\}\rightarrow\{1111\}$.

**Equações de equilíbrio:**

A seguir, apresentamos uma breve discussão de algumas das equações de transição adicionais entre os estados do sistema.

(i) Quando há apenas 1 servidor ocupado no sistema:
Por exemplo, para o estado {0100}, em que apenas o servidor 2 está ocupado temos:

$$(\lambda + \mu_2).P_{\{0100\}} = \lambda_{1a}^{[1]}.P_{\{0000\}} + \mu_1.P_{\{1100\}} + \mu_3.P_{\{0110\}} + \mu_4.P_{\{0101\}} \quad (4.57)$$

Note no lado esquerdo da expressão (4.57), que no estado {0100} não ocorre perda de chamadas no sistema. Dado que o servidor preferencial (servidor 2) está ocupado, chamadas do tipo 1 e 2a no átomo 1a e do tipo 2a no átomo 2a são atendidas pelo servidor 3, e a transição é $\{0100\}\rightarrow\{0110\}$. Além disso, chamadas do tipo 2b e 3 nos átomos 1b e 2b são atendidas pelos servidores 1 e 3 ($\{0100\}\rightarrow\{1110\}$), e chamadas do tipo 3 no átomo 3a são atendidas somente pelos servidores 3 e 4 ($\{0100\}\rightarrow\{0111\}$). No lado direito da expressão (4.57), notamos que a transição $\{0000\}\rightarrow\{0100\}$ ocorre com a chegada de uma chamada tipo 1 no átomo 1a atendida pelo servidor preferencial 2 (tabela 4.8).

(ii) Quando há dois servidores ocupados no sistema:

Por exemplo, para o estado {0110}, que corresponde ao estado em que os servidores 2 e 3 estão ocupados e 1 e 4 livres, temos a seguinte equação de equilíbrio:

$$\begin{aligned}&((\lambda - \lambda_{1a}^{[1]} - \lambda_{1a}^{[2a]} - \lambda_{2a}^{[1]} - \lambda_{2a}^{[2a]}) + \mu_2 + \mu_3).P_{\{0110\}} = (\lambda_{1a}^{[2a]} + \lambda_{2a}^{[2a]}).P_{\{0000\}} + (\lambda_{1a}^{[1]} + \lambda_{1a}^{[2a]} + \\ &\lambda_{2a}^{[1]} + \lambda_{2a}^{[2a]}).P_{\{0100\}} + (\lambda_{1a}^{[1]} + \lambda_{1a}^{[2a]} + \lambda_{2a}^{[1]} + \lambda_{2a}^{[2a]}).P_{\{0010\}} + \mu_1.P_{\{1110\}} + \mu_4.P_{\{0111\}}\end{aligned} \quad (4.58)$$

Observe no lado esquerdo da expressão (4.58) que no estado {0110}, chamadas do tipo 1 e 2a nos átomos 1a e 2a são perdidas para o sistema, pois seu servidor preferencial e *backup* estão ocupados (veja tabela 4.8). No lado direito da equação, notamos, por exemplo, que a transição $\{0000\}\rightarrow\{0110\}$ ocorre com a chegada de chamadas do tipo 2a nos átomos 1a e 2a.

(iii) Quando há três servidores ocupados no sistema:

Analisando, por exemplo, o estado {1110}, temos a equação de equilíbrio:

$$(\lambda_{3a}^{[1]} + \lambda_{3a}^{[2a]} + \lambda_{3b}^{[2b]} + \lambda_{3a}^{[3]} + \mu_1 + \mu_2 + \mu_3).P_{\{1110\}} = (\lambda_{1b}^{[3]} + \lambda_{2b}^{[3]}).P_{\{0000\}} + (\lambda_{1a}^{[1]} + \lambda_{1a}^{[2a]} + \lambda_{1b}^{[2b]} + \lambda_{1b}^{[3]} + \lambda_{2a}^{[1]} + \lambda_{2a}^{[2a]} + \lambda_{2b}^{[2b]} + \lambda_{2b}^{[3]}).P_{\{1000\}} + (\lambda_{1b}^{[2b]} + \lambda_{1b}^{[3]} + \lambda_{2b}^{[2b]} + \lambda_{2b}^{[3]}).P_{\{0100\}} + (\lambda_{1b}^{[2b]} + \lambda_{1b}^{[3]} + \lambda_{2b}^{[2b]} + \lambda_{2b}^{[3]}).P_{\{0010\}} + (\lambda_{1a}^{[1]} + \lambda_{1a}^{[2a]} + \lambda_{1b}^{[2b]} + \lambda_{1b}^{[3]} + \lambda_{2a}^{[1]} + \lambda_{2a}^{[2a]} + \lambda_{2b}^{[2b]} + \lambda_{2b}^{[3]}).P_{\{1100\}} + (\lambda_{1a}^{[1]} + \lambda_{1a}^{[2a]} + \lambda_{1b}^{[2b]} + \lambda_{1b}^{[3]} + \lambda_{2a}^{[1]} + \lambda_{2a}^{[2a]} + \lambda_{2b}^{[2b]} + \lambda_{2b}^{[3]}).P_{\{1010\}} + (\lambda_{1b}^{[2b]} + \lambda_{1b}^{[3]} + \lambda_{2b}^{[2b]} + \lambda_{2b}^{[3]}).P_{\{0110\}} + \mu_4.P_{\{1111\}}$$

(4.59)

Note no lado esquerdo da equação (4.59) que somente chamadas geradas nos átomos 3a e 3b podem ser atendidas pelo sistema, quando no estado {1110}, pois somente o servidor 4 está livre e atende chamadas sozinho, possivelmente com a ajuda de servidores de outros SAEs mais próximos.

No lado direito da equação, verificamos que três servidores (1, 2 e 3) podem passar de livre a ocupados simultaneamente quando chamadas do tipo 3 são geradas nos átomos 1 e 2 (transição $\{0000\} \rightarrow \{1110\}$). Além disso, note que as transições $\{0100\} \rightarrow \{1110\}$ e $\{0010\} \rightarrow \{1110\}$ podem ocorrer com a chegada de chamadas do tipo 2b nos átomos 1 e 2, dado que se um dos dois servidores preferenciais estiver ocupado, o terceiro servidor também é despachado. Por outro lado, chamadas dos tipos 2a, 2b e 3 que encontram um único servidor disponível são atendidas por este servidor, por exemplo, nas transições: $\{0110\} \rightarrow \{1110\}$, $\{1010\} \rightarrow \{1110\}$ e $\{1100\} \rightarrow \{1110\}$.

(iv) Quando todos os servidores estão ocupados no sistema:

A equação de equilíbrio no estado {1111} é:

$$(\mu_1 + \mu_2 + \mu_3 + \mu_4).P_{\{1111\}} = (\lambda_{1b}^{[3]} + \lambda_{2b}^{[3]}).P_{\{0001\}} + (\lambda_{3a}^{[3]}).P_{\{1000\}} + (\lambda_{1b}^{[2b]} + \lambda_{1b}^{[3]} + \lambda_{2b}^{[2b]} + \lambda_{2b}^{[3]}).P_{\{0011\}} + (\lambda_{1b}^{[2b]} + \lambda_{1b}^{[3]} + \lambda_{2b}^{[2b]} + \lambda_{2b}^{[3]} + \lambda_{3b}^{[2b]}).P_{\{0101\}} + (\lambda_{3b}^{[2b]}).P_{\{0110\}} + (\lambda_{1a}^{[2a]} + \lambda_{1b}^{[2b]} + \lambda_{1b}^{[3]} + \lambda_{2a}^{[2a]} + \lambda_{2b}^{[2b]} + \lambda_{2b}^{[3]} + \lambda_{3a}^{[3]}).P_{\{1001\}} + (\lambda_{3a}^{[3]}).P_{\{1010\}} + (\lambda_{3a}^{[2a]} + \lambda_{3b}^{[2b]} + \lambda_{3a}^{[3]}).P_{\{1100\}} + (\lambda_{1b}^{[2b]} + \lambda_{1b}^{[3]} + \lambda_{2b}^{[2b]} + \lambda_{2b}^{[3]} + \lambda_{3b}^{[2b]}).P_{\{0111\}} + (\lambda_{1a}^{[1]} + \lambda_{1a}^{[2a]} + \lambda_{1b}^{[2b]} + \lambda_{1b}^{[3]} + \lambda_{2a}^{[1]} + \lambda_{2a}^{[2a]} + \lambda_{2b}^{[2b]} + \lambda_{2b}^{[3]} + \lambda_{3a}^{[3]}).P_{\{1011\}} + (\lambda).P_{\{1101\}} + (\lambda_{3a}^{[1]} + \lambda_{3a}^{[2a]} + \lambda_{3b}^{[2b]} + \lambda_{3a}^{[3]}).P_{\{1110\}}$$

(4.60)

As transições da expressão acima são resultantes da política de despacho descrita na tabela 4.8, de acordo com os possíveis servidores (preferencial e servidores *backup*) da lista de preferência de despacho.

**Medidas de Desempenho:**

Carga de trabalho dos servidores

Como nos exemplos 2, 3, 4 e 5 anteriores, a carga de trabalho de cada servidor $\rho_j$ é calculada considerando todos os estados em que o mesmo se encontra ocupado. Por exemplo, a carga de trabalho do servidor 2 no sistema do exemplo de quatro servidores acima é calculada da seguinte forma:

$$\rho_2 = P_{0100} + P_{1100} + P_{0110} + P_{0101} + P_{1110} + P_{0111} + P_{1101} + P_{1111}$$

Probabilidade de Perda

A probabilidade de perda pode ser calculada como nos exemplos anteriores. De forma similar ao exemplo 4, podemos também calcular a probabilidade de perda para qualquer chamada do sistema e para cada tipo de chamada. No exemplo 6, podemos ter: probabilidade de perda de chamadas tipo 1 ($P_p^{[1]}$), probabilidade de perda de chamadas tipo 2a ($P_p^{[2a]}$), probabilidade de perda de chamadas tipo 2b ($P_p^{[2b]}$), probabilidade de perda de chamadas tipo 3 ($P_p^{[3]}$) e prob1ailidade de perda para qualquer chamada do sistema ($P_p$). Por exemplo, a medida $P_p$ é calculada da seguinte forma:

$$P_p = \sum_B \left( \frac{\sum_{i \in N_A^{[b]}} \sum_{m=1}^{4} (\lambda_i^{[m]})}{\lambda} P_B \right), \qquad (4.62)$$

Onde $N_A^{[b]}$ corresponde ao conjunto de átomos com todos os servidores de sua lista de preferência ocupados no estado *B* e *m* = 1, 2a, 2b e 3 (tipo de chamada).

Lembre-se que, uma chamada tipo 1 ou tipo 2a não é atendida pelo sistema se seus dois servidores da lista de preferência estão ocupados. Por exemplo, no estado 0110, chamadas 1 e

2a geradas nos átomos 1 ou 2 são perdidas para o sistema, pois o servidor 3 não pode atendê-las. Por outro lado, chamadas do tipo 2b ou 3 são perdidas somente se os três servidores da lista de preferência estiverem ocupados. Por exemplo, no estado 1110, chamadas tipo 2b e 3 nos átomos 1 e 2 não são atendidas.

Freqüências de despacho

Como no exemplo 2 e 4, podemos calcular freqüências de despacho do servidor *j* ao átomo *i* para cada tipo de chamada *m* = 1, 2a, 2b e 3.

Por exemplo, para chamadas do tipo 1, podemos calcular $f_{ji}^{[1]}$ e $f'^{[1]}_{ji}$ utilizando as expressões 4.25 e 4.33, respectivamente. As medidas de freqüência relacionadas a duplo despacho podem ser obtidas, de forma similar ao exemplo 4. Por exemplo, as freqüências de duplo e triplo despacho considerando todos os despachos do sistema podem ser definidas para chamadas tipo 1, 2a, 2b e 3 da seguinte forma:
(i) Chamadas tipo 2a e 2b:

$$f'^{[2a]}_{(j,k)i} = \frac{\frac{\lambda_i^{[2a]}}{\lambda}\sum_{B\in E_{ji}} P_B}{(1-P_p)} \text{ e } f'^{[2b]}_{(j,k)i} = \frac{\frac{\lambda_i^{[2b]}}{\lambda}\sum_{B\in E_{(j,k)i}} P_B}{(1-P_p)}, \tag{4.63}$$

Note que, diferentemente dos atendimentos do tipo 2a, no caso de chamadas tipo 2b, o servidor *j* ou *k* pode ser o terceiro servidor da lista de preferência de despacho do átomo *i*, se o primeiro ou segundo servidor estiver ocupado.

(ii) Chamadas tipo 3 (atendidas por dois servidores *j* e *k* - únicos disponíveis na lista de preferência de despacho):

$$f'^{[3]}_{(j,k)i} = \frac{\frac{\lambda_i^{[3]}}{\lambda}\sum_{B\in F(j,k)i} P_B}{(1-P_p)}, \tag{4.64}$$

onde $F_{(j,k)i}$ corresponde ao conjunto de estados em que os servidores *j* e *k* são os únicos servidores disponíveis na lista de preferência de despacho do átomo *i*.

Diversas medidas adicionais de freqüência podem ser definidas tais como: freqüências de único despacho para os 4 tipos de chamadas, freqüências de triplo despacho para chamadas

tipo 3 (levando em considerando todos os despachos do sistema ou todos os despachos tipo 3), freqüências de único despacho atendidas por servidor *backup*; freqüências de duplo despacho para chamadas do tipo 2b e 3 nas quais os servidores *j* e *k* são: o primeiro e segundo, primeiro e o terceiro ou o segundo e o terceiro da lista de preferência de despacho; entre outras medidas.

Medidas de tempo de viagem adicionais:

Além das medidas agregadas de tempo de viagem descritas na seção 4.2, podemos definir outras medidas adicionais de acordo com o tipo de chamada e o número de servidores enviados (único, duplo ou triplo despacho). Entre estas estão: tempo médio de viagem no sistema para chamadas tipo *m* quando ocorre único despacho, tempo médio de viagem para chamadas com duplo despacho para chamadas tipos 2a e 2b e tipo 3 (quando somente dois servidores estão disponíveis), tempo médio de viagem para a primeira, segunda e terceira (no caso de triplo despacho) ambulância a chegar no local da chamada, entre outras.

Uma extensão interessante do presente modelo Hipercubo é considerar chamadas atendidas na base (i.e., chamadas tipo 1a), por meio de um terceiro estado, similarmente ao modelo Hipercubo da seção 4.3. Como discutido naquela seção, o espaço de estados cresce de $O(2^N)$ para $O(3^N)$.

# 5. Modelos Prescritivos e Abordagem combinando modelo Hipercubo com um Algoritmo Genético:

Neste capítulo inicialmente alguns modelos prescritivos de localização probabilística e configuração do sistema encontrados na literatura são brevemente revisados. Em seguida, discute-se a metodologia de Algoritmo Genético, e se propõe uma abordagem combinando o modelo Hipercubo com um algoritmo genético para otimizar a configuração do sistema.

Os modelos prescritivos são modelos que, uma vez resolvidos por meio de técnicas de otimização, indicam boas (senão ótimas) alternativas de operação do sistema. A maioria dos modelos prescritivos utilizados na análise dos SAEs são os modelos de localização de instalações (por exemplo, hospitais, estações de bombeiro, delegacias de polícia, garagens satélites, etc) .

Os primeiros modelos propostos na literatura para tratar problemas de localização em sistemas de emergência foram modelos determinísticos. Em geral, estes modelos representam o problema num grafo, e técnicas de teoria de grafos, otimização combinatória e heurísticas são aplicadas para resolvê-los.

Os modelos determinísticos com restrição de cobertura têm sido largamente empregados para os problemas de localização relacionados aos SAEs. Uma área é considerada coberta se está a menos da distância crítica (pré-determinada de acordo com o nível de serviço desejado) de pelo menos uma das instalações (ou servidor), independentemente da mesma se encontrar não disponível quando solicitada (CHIYOSHI et al, 2000).

Um dos primeiros modelos com restrição de cobertura para sistemas de emergência foi o Problema de Localização de Cobertura de Conjuntos (SCLP – *Set Covering Location Problem*), estudado por TOREGAS et al (1971). O modelo busca encontrar o mínimo número de servidores necessários para oferecer cobertura a toda população da região considerada. O modelo apresenta a desvantagem de que, quanto melhor o nível de serviço oferecido, o número de servidores necessários para oferecer cobertura total pode ser inviável diante das restrições de recursos disponíveis.

Um problema derivado do SCLP é o Problema de Localização dos *p*-centros (*p-center Location Problem)*, cujo objetivo é localizar *p* instalações de forma a minimizar a máxima distância de um dado nó de demanda à sua instalação mais próxima. O método de solução proposto por MINIEKA (1970) e CHRISTOFIDES & VIOLA (1971) baseia-se em solucionar uma seqüência de problemas de cobertura de conjuntos (SCLP), reduzindo-se sucessivamente a distância crítica a cada iteração. O processo prossegue até que a redução na distância crítica requeira que $p+1$ instalações sejam localizadas.

Como alternativa ao SCLP, CHURCH & REVELLE (1974) propuseram o Problema de Localização de Máxima Cobertura (MCLP – *Maximal Covering Location Problem*), que não supõe que toda população deva ser coberta e considera a quantidade de demanda em cada nó na formulação. Dada uma distância crítica, o problema busca localizar um número *p* de instalações de forma a maximizar a população coberta.

A idéia do MCLP foi adotada por SCHILLING et al (1979) em um modelo para localização de estações de bombeiros. O problema chamado FLEET (*Facility Location Equipment Emplacement Technique*), busca por uma ideal localização das estações de bombeiro, assim como a melhor alocação das viaturas (carros pipas e carros resgate) em cada estação, de forma a maximizar a população coberta com base em uma distância crítica de cada tipo de viatura.

EATON et al (1985) utilizaram o modelo MCLP, para estudar o sistema de atendimento médico emergencial em Austin, Texas. Neste estudo, foram considerados dois tipos de chamadas: chamadas consideradas de alta prioridade, que requerem suporte avançado de tratamento e chamadas com menor prioridade, que requerem suporte básico de tratamento. O MCLP possibilitou vários tipos de análises, como por exemplo, variação no número de cada tipo de veículo no sistema, variação do tempo crítico de cobertura, e variação dos locais candidatos para localização dos servidores. A implementação de um plano proposto por aquele estudo resultou em significante redução de custos e redução no tempo médio de resposta, apesar do aumento da taxa de demanda no sistema.

No estudo de KARASAKAL & KARASAKAL (2004), os autores propuseram o modelo MCLP-P, como uma extensão do MCLP que considera cobertura parcial de um nó de demanda. Desta forma, o modelo estabelece que o nó de demanda pode estar totalmente coberto dentro de uma distância crítica mínima $S_1$, parcialmente coberto dentro de uma

distância crítica máxima $S_2$ e não coberto se a instalação mais próxima estiver localizada fora da distância crítica máxima.

Estes são apenas alguns exemplos de estudos propondo modelos determinísticos para localização em SAEs. Os principais modelos determinísticos estão descritos com mais detalhes em KOLESAR & SWERSEY (1986), REVELLE (1989), LOUVEAUX (1993), SWERSEY (1994), DASKIN (1995), OWEN & DASKIN (1998), CHIYOSHI et al. (2000) e BROTCORNE et al. (2003).

A principal desvantagem destes modelos é que não consideram a possibilidade de os servidores estarem ocupados quando ocorre um chamado. Conforme SWERSEY (1994), tais modelos são mais adequados para analisar SAEs que apresentam índices muito altos de ociosidade, o que permite admitir que há sempre um servidor disponível e desprezar o caráter estocástico do problema.

Os modelos determinísticos com enfoque em cobertura adicional constituem uma segunda classe de modelos determínísticos que trouxeram importantes contribuições para considerar a indisponibilidade dos servidores. Estes modelos foram propostos para estudar os sistemas congestionados onde há significante probabilidade de uma chamada, gerada em uma certa região, encontre o primeiro (muitas vezes o único) servidor cobrindo esta região, não disponível para atendê-la.

Modelos de cobertura adicional aparecem em HOGAN & REVELLE (1986). Baseando-se no SCLP, os autores propuseram um modelo que assegura a primeira cobertura a toda população e ao mesmo tempo, procura maximizar a população coberta por ao menos um servidor adicional. O modelo é conhecido como BACOP I (*Backup Coverage Problem*). O BACOP II foi proposto como uma extensão do MCLP. O problema visa oferecer múltipla cobertura às áreas de maior demanda (nas quais há maior chance de dois nós concorrerem por um servidor), antes de assegurar primeira cobertura às áreas de menor demanda. Este problema é formulado com duas funções objetivo, a primeira maximiza a população com primeira cobertura (função objetivo do MCLP) e a segunda maximiza a população com duas ou mais coberturas, considerando a quantidade de demanda em cada nó.

## 5.1 Modelos Probabilísticos

Embora os modelos prescritivos com cobertura adicional tenham significado um importante avanço em considerar a indisponibilidade dos servidores, em muitos casos este tipo de tratamento é insuficiente. Como discutido no capítulo 2, os sistemas de emergência, em sua maioria, caracterizam-se pelo alto grau de aleatoriedade relacionada não só a disponibilidade dos servidores, como também ao processo de resposta e demanda por serviço. Os principais modelos de localização probabilísticos são voltados para considerar a aleatoriedade relacionada à disponibilidade dos servidores. A seguir, apresentamos brevemente alguns modelos, entre os mais relevantes que aparecem na literatura.

### Problema de Localização de Máxima Cobertura Esperada (MEXCLP – *Maximum Expected Covering Location Problem*)

O estudo de DASKIN (1983) foi um dos trabalhos pioneiros em estender os modelos determinísticos de localização para análise dos sistemas de emergência, considerando a possibilidade de os servidores estarem indisponíveis devido à congestão. O modelo desenvolvido é uma extensão probabilística do problema de localização de máxima cobertura (MCLP) proposto por CHURCH & REVELLE (1974), e é chamado de problema de localização de máxima cobertura esperada (MEXCLP - *Maximum Expected Covering Location Problem*). O MEXCLP busca maximizar a cobertura esperada, dado que há uma certa probabilidade de os servidores estarem ocupados quando ocorre um chamado. As principais hipóteses do modelo são: os servidores operam independentemente entre si, e as probabilidades de que eles estejam ocupados são idênticas e não dependem da localização dos servidores. O modelo também admite que mais de um servidor pode estar localizado em um mesmo nó se necessário.

No trabalho de TAVAKOLI & LIGHTNER (2004), os autores propõem combinar o modelo FLEET de SCHILLING et al (1979) com o MEXCLP. O modelo é conhecido como MOFLEET (*Multiple Coverage One Unit FLEET*) e procura minimizar a população não coberta, de forma que, os nós com maior demanda possam ser atendidos por até M servidores localizados a menos de uma distância crítica (múltipla cobertura). O modelo foi aplicado ao sistema médico emergencial de Fayetteville, Carolina do Norte.

**Problema de Localização de Máxima Disponibilidade (MALP – *Maximum Availability Location Problem*)**

Alternativamente ao MEXCLP, REVELLE & HOGAN (1989), propuseram o problema de localização de máxima disponibilidade (MALP – *Maximum Availability Location Problem*). O MALP também é derivado do MCLP, e o método consta de maximizar a população coberta com um servidor disponível dentro de um tempo de resposta estabelecido, considerando um fator $\alpha$ de confiabilidade. Como no MEXCLP, o MALP admite que os servidores operam de forma independente, com mesma taxa de ocupação, a qual independe da localização dos mesmos. Os autores apresentam duas versões: o MALP I considera que a probabilidade de ocupação de todos os servidores é a mesma, e o MALP II introduz uma fórmula para estimar as diferentes probabilidades de ocupação entre os servidores, considerando que a área primaria está isolada do resto da região.

Os modelos MEXCLP e MALP serviram de base para outros recentes estudos em problemas de localização. Uma descrição mais detalhada sobre a formulação dos modelos MEXCLP e MALP é encontrada nos Anexos deste texto.

## 5.2 Modelos de Localização probabilísticos que incorporam Teoria de Filas:

Os modelos MEXCLP e MALP são modelos prescritivos que consideram a aleatoriedade associada à disponibilidade dos servidores do sistema. No entanto, estes modelos assumem certas simplificações que na maioria dos casos não se aplicam a sistemas de atendimento emergencial. São elas :

(i) independência dos servidores. Nos SAEs os servidores não operam de forma independente, pois podem cooperar entre si como forma de reduzir o tempo médio de resposta;

(ii) taxa de ocupação idêntica para todos os servidores. Como nos SAEs os servidores estão espacialmente distribuídos, a taxa de ocupação varia entre eles;

(iii) taxa de ocupação não varia com a localização dos servidores. Nos SAEs, a taxa de ocupação dos servidores depende da localização dos mesmos e da política de despacho adotada.

Além disso, outras características aleatórias do sistema relacionadas aos processos de chegada e atendimento são importantes na análise dos sistemas de atendimento emergencial. Estas variáveis podem ser tratadas incorporando modelos de Teoria de Filas aos modelos prescritivos. Vários estudos nesta direção têm sido desenvolvidos desde do início da década de 70. Algumas referências importantes podem ser encontradas em SWERSEY (1994), CHIYOSHI et al. (2000) e BROTCORNE et al. (2003). Um exemplo é o estudo de MARIANOV & REVELLE (1996) que estende o MALP introduzindo o modelo Q-MALP, no qual os tempos de viagem e a distância entre os nós são tratados como variáveis aleatórias e as técnicas de Teoria de Filas são empregadas para avaliar a fração de ocupação de cada região no MALP.

Como mencionado anteriormente, o modelo Hipercubo de LARSON (1974) mostra-se como o modelo de filas mais adequado para tratar os fatores probabilísticos dos SAEs relacionados, principalmente, à distribuição temporal e espacial dos servidores. Como este modelo é baseado em teoria de filas espacialmente distribuídas, ele permite descrever o sistema levando em conta a individualidade de seus servidores e a cooperação e/ou interação entre eles. A seguir, apresentamos apenas os modelos prescritivos probabilísticos que incorporam o modelo Hipercubo de filas, dado que este é o foco do presente estudo.

**Problema de Localização de Máxima Cobertura Esperada Ajustado (AMEXCLP - *Adjusted Maximal Expected Covering Location Problem*)**

BATTA et al. (1989) revisitaram o modelo MEXCLP e propuseram um procedimento iterativo integrado ao modelo Hipercubo como forma de relaxar as três hipóteses simplificadoras do MEXCLP, ou seja: (i) os servidores operam independentemente; (ii) os servidores têm a mesma taxa de ocupação e (iii) a taxa de ocupação dos servidores não varia com suas localizações.

Este procedimento procura localizar servidores numa região representada por uma rede, de maneira a maximizar a cobertura total esperada do sistema. Cada vértice da rede corresponde ao centróide de um átomo, ou seja, uma possível localização para um servidor. O método de solução proposto é um procedimento iterativo baseado na heurística de substituição de vértices de TEIZ & BART (1968), onde em cada iteração, o modelo Hipercubo é resolvido para estimar a cobertura esperada da configuração corrente. Basicamente, os passos deste

procedimento iterativo resumem-se em: (a) inicie com uma solução candidata para localização dos servidores, e utilize o modelo Hipercubo para calcular a cobertura esperada; (b) considere todo conjunto de localizações obtido pela substituição por um único nó e, para cada um, calcula-se a cobertura esperada através do modelo Hipercubo; (c) a solução que possui a maior cobertura esperada torna-se a nova solução corrente e o processo iterativo é repetido, até não ser mais possível melhorá-la pelo procedimento de substituição por um único nó.

Como a cobertura esperada não é uma medida de desempenho obtida diretamente pelo modelo Hipercubo, BATTA et al (1989) formularam uma expressão para calculá-la a partir dos resultados do Hipercubo. Os autores também formularam um novo modelo, como extensão do MEXCLP. Para relaxar a hipótese de independência entre os servidos, foram introduzidos os fatores de correção utilizados por LARSON (1975) no modelo Hipercubo aproximado. O novo modelo foi chamado Problema de Localização de Máxima Cobertura Esperada Ajustado (AMEXCL - *Adjusted Maximal Expected Covering Location Problem*). A formulação do AMEXCL também é apresentada nos Anexos deste texto.

SAYDAM et al (1994) avaliaram a acuracidade dos resultados obtidos pelo AMEXCLP, e mostraram que o modelo obtém uma solução "ótima" ou "perto da ótima" para o conjunto de localizações, mas pode resultar em uma significante sobre ou sub estimativa de cobertura. Os autores sustentam a recomendação de BATTA et al. (1989) no uso do modelo Hipercubo com modelos de otimização.

CHIYOSHI et al (2003) investigaram as diferenças obtidas no cálculo da cobertura esperada utilizando os modelos MEXCLP, AMEXCLP e o procedimento iterativo incorporando o modelo Hipercubo, proposto por BATTA et al (1989), o qual eles chamaram de HLM (*Hypercube Location Model*). O estudo mostra que, como os modelos MEXCLP e AMEXCLP não podem considerar a possibilidade das chamadas em fila, estes três métodos não podem ser diretamente comparados considerando sistemas com fila de capacidade infinita. Uma comparação mais adequada é comparar os três métodos para o cálculo da cobertura esperada em um sistema sem filas. No entanto os autores sustentam que o HLM apresenta-se como o método mais promissor, dado que pode computar as chamadas em fila no cálculo da cobertura esperada.

**Problema de Localização Não Linear de Máxima Cobertura Experada (NMEXCLP - *Non Linear Maximal Expected Covering Location Problem*)**

SAYDAM & MACKNEW (1985) propuseram uma extensão do MEXCLP de DASKIN (1983) utilizando uma aproximação de programação separável. Os autores consideram que a localização de instalações em um sistema de emergência implica:

- localizar um número suficiente de instalações para satisfazer a demanda;
- distribuí-las geograficamente para que possam cobrir a demanda dentro do padrão de resposta, considerando que estes padrões são diferentes para áreas urbanas e rurais.

O modelo tem a flexibilidade para determinar esquemas ótimos de localização, de forma a assegurar múltipla cobertura às áreas de demanda, considerando diferentes tipos de restrições e vários critérios de resposta simultânea. Cada nó de demanda tem um termo não-linear na função objetivo.

**Problema de Localização de Máxima Disponibilidade Estendido (EMALP – *Extended Maximal Availability Location Problem*)**

No estudo de RIVAS (2002) e GALVAO et al (2003), um método heurístico foi proposto como uma extensão do MALP desenvolvido por REVELLE & HOGAN (1989). Seguindo a idéia de BATTA et al (1989), os autores implementaram um procedimento de substituição de vértice que utiliza o modelo Hipercubo para avaliar a configuração do sistema a cada iteração. O modelo, chamado de modelo de localização de máxima disponibilidade estendido (EMALP), também utiliza os fatores de correção de LARSON (1975) para relaxar as hipóteses de independência entre os servidores admitida no MALP e permite o cálculo das probabilidades de ocupação de cada servidor através do modelo Hipercubo.

A formulação do MALP foi modificada com a introdução de uma nova variável e dos fatores de correção. Esta variável é necessária, pois para considerar cada servidor individualmente não é apenas necessário saber se há um servidor no nó de demanda $j$, mas também é preciso identificar qual servidor está localizado neste nó.

## 5.3 O uso de métodos baseados em meta-heurísticas para tratar os problemas de localização nos SAEs.

Estudos que utilizam meta-heurísticas para solucionar problemas de otimização têm sido cada vez mais explorados pelos analistas em Pesquisa Operacional. Além da relativa facilidade de implementação, estes métodos podem prescrever, em razoável tempo computacional, a solução ótima ou perto da ótima para complexos problemas combinatórios. No caso dos problemas voltados para os sistemas de atendimento emergencial, importantes contribuições vêm sendo obtidas em estudos recentes.

GENDREAU et al (1997) aplicaram *Busca Tabu* para um modelo de localização de dupla cobertura que formularam, o DSM (*Double Standard Model*). Neste modelo, os autores estabelecem dois padrões de cobertura $r_1$ e $r_2$, onde $r_1 < r_2$, sendo que toda demanda deve ser coberta por uma instalação localizada dentro de $r_2$ e ao menos uma proporção $\beta$ da demanda deve coberta por uma instalação dentro de $r_1$. O objetivo é maximizar a população com dupla cobertura. O valor de $r_1$ é atribuído de acordo com as especificações do *United States Emergency Medical Services Act* de 1973. O estudo utilizou instâncias geradas aleatoriamente e dados obtidos da cidade de Montreal, Canadá. Os autores mostraram que o algoritmo com *Busca Tabu* é uma técnica eficiente para solucionar o problema, produzindo soluções perto da ótima em tempo computacional relativamente pequeno.

Em GRENDEAU et al (2001), os autores também aplicaram *Busca Tabu* para um modelo de localização dinâmico. Os modelos dinâmicos de localização têm o objetivo de determinar novas políticas para despachar as ambulâncias em cada instante, utilizando as mais recentes informações disponíveis sobre o sistema.

COSTA (2004) propôs um método que utiliza Algoritmos Genéticos (GA) para determinar as zonas de atendimento das unidades de serviços emergenciais do corpo de bombeiros em Curitiba, PR., o sistema SIATE (Serviço Integrado de Atendimento ao Trauma em Emergência). A abordagem proposta consta em dividir o sistema em N zonas de atendimento, e localizar cada ambulância em uma zona de forma a minimizar uma medida de desempenho estabelecida. Para isso associa-se ao algoritmo genético um procedimento de atendimento simulado (método determinístico e dinâmico) proposto para avaliação da *fitness* das soluções

geradas. A melhor configuração final obtida pelo algoritmo genético é também avaliada pelo modelo Hipercubo de LARSON (1974) para sistemas com fila de capacidade infinita.

GALVAO et al. (2005) utilizam *Simulated Annealing* (SA) para resolver o AMEXCLP de BATTA et al (1989) e o EMALP proposto por GALVAO et al (2003). Como mencionado anteriormente, estes modelos foram inicialmente resolvidos através da inserção do método aproximado do modelo Hipercubo de LARSON (1975) em uma heurística de substituição de vértice. Em GALVAO et al. (2005), os autores também inserem o modelo Hipercubo aproximado em um algoritmo de *Simulated Annealing* e a metodologia de SA utilizada é a mesma descrita em CHIYOSHI & GALVAO (2000). Os resultados obtidos pelo algoritmo SA são comparados com os resultados obtidos pela heurística de substituição de vértices (VS), para o mesmo conjunto de problemas disponíveis na literatura. O estudo mostrou que o algoritmo SA apresenta melhor desempenho em termos da qualidade da solução obtida e da capacidade de encontrar a melhor solução. No entanto, as diferenças dos resultados do SA para o VS no caso do AMEXCLP são de pouca aplicação prática, enquanto que no caso do EMALP estas diferenças são significantes.

SAYDAM & AYTUG (2003) desenvolveram um Algoritmo Genético (GA) para o MEXCLP combinado com o algoritmo aproximado do modelo Hipercubo modificado por JARVIS (1985). A função *fitness* utilizada neste algoritmo é a função objetivo do MEXCLP, modificada para incluir as probabilidades de ocupação de cada servidor $\rho_j$, calculadas pelo método aproximado de Jarvis. Esta função *fitness* passa a ser:

$$\sum_{i=1} h_i \left( 1 - \prod_{j \in K_i} \rho_j \right), \tag{5.1}$$

onde $K_i$ é o conjunto de servidores que oferecem cobertura ao nó de demanda $i$ e $h_i$ é a demanda no átomo $i$.

Desta forma, em cada iteração, o modelo Hipercubo é calculado para toda a população de soluções (avaliadas no ciclo genético). Os resultados obtidos foram satisfatórios do ponto de vista da qualidade das soluções obtidas, melhorando significantemente os resultados obtidos pelo MEXCLP. Além de enfatizar a eficiência e acuracidade do algoritmo, os autores também

incentivam o uso da abordagem GA/ Hipercubo nos estudos com enfoque no balanceamento das cargas de trabalho entre os servidores.

A abordagem GA/Hipercubo proposta em SAYDAM & AYTUG (2003), também serviu de base para o presente estudo, assim como os trabalhos de CHIYOSHI et al. (2003) e GALVÃO et al. (2005). A seguir, discutimos brevemente a idéia básica dos Algoritmos Genéticos e suas aplicações em problemas de localização, para em seguida propor uma abordagem GA/Hipercubo para tratar SAEs em rodovias.

## 5.4 Algoritmos Genéticos e suas aplicações em problemas de localização:

Algoritmos Genéticos (*Genetic Algorithms* - GAs), originalmente desenvolvidos por HOLLAND (1975), procuram simular o processo biológico de evolução dos organismos na natureza. Por este processo, uma população de indivíduos evolui de acordo com os princípios da seleção natural. Desta forma, os indivíduos mais adaptados ao seu ambiente devem ter maiores chances de sobreviver e procriar, ao passo que, indivíduos menos aptos devem ser naturalmente extintos. Após gerações sucessivas, indivíduos mais adaptados irão crescer em número na população, pois a combinação de "bom" material genético deve produzir melhores descendentes. Além do processo de reprodução (*crossover*), o processo de mutação de indivíduos selecionados pode contribuir para a geração da população de descendentes.

Seguindo a idéia da adaptação das espécies na natureza, um GA parte de uma população inicial de indivíduos (também chamados cromossomos). Cada cromossomo representa uma solução possível do problema a ser estudado. A cada geração (iteração) da busca, o algoritmo avalia e seleciona os melhores indivíduos para gerar a próxima população de acordo com o valor da *"fitness"*, ou seja, um valor relacionado à função objetivo do problema analisado. Através deste procedimento, o algoritmo tenta produzir descendentes melhores que seus pais. Após o procedimento de seleção, os cromossomos escolhidos podem permanecer intactos na próxima população ou reproduzir, transmitindo a sua herança genética para gerar melhores indivíduos. A reprodução consiste da recombinação de dois cromossomos através do procedimento de *crossover* (cruzamento), possivelmente seguido pelo processo de mutação. O ciclo se repete e espera-se que após certo número de gerações sucessivas, a solução tende a ficar perto da solução ótima do problema estudado.

Muitos trabalhos descrevem os GAs e outros Algoritmos Evolucionários em detalhes, por exemplo, GOLDBERG (1989), MICHALEWICZ (1996), REEVES (1997) e BEASLEY (2000). O processo de busca de um GA pode ser resumido da seguinte forma:

*Crie uma população inicial de cromossomos;*
*Enquanto o critério de parada não é satisfeito:*
*Selecione os cromossomos pais;*
*Se crossover então realize crossover;*
*Se mutação então realize mutação;*
*Avalie os cromossomos filhos;*
*Determine qual a melhor solução encontrada através das gerações.*

As principais características que distinguem o GA das demais heurísticas são mencionadas, por exemplo, em GLOVER et al (1995), MICHALEWICZ (1996), entre elas:

- GA realiza a busca considerando uma população de soluções a cada iteração, ao invés de uma única solução. Conseqüentemente, há menos chance da busca ser bloqueada em pontos de ótimo local;
- GA pode ser representado por códigos dos parâmetros de interesse, ao invés dos próprios parâmetros;
- GA apresenta natureza probabilística, e a troca de informações entre diferentes soluções e gerações é realizada de acordo com regras aleatórias.

Várias referências de aplicações de GA na área de otimização são citadas por REEVES (1997). Por exemplo, problemas relacionados à localização, sequenciamento, árvore de *Steiner*, roteirização de veículos, *layout* e outros.

Os principais componentes para a aplicação de GA são:

1. Representação adequada dos cromossomos (soluções);
2. Operadores efetivos de *crossover* e mutação;
3. Uma rotina para criar a população inicial;
4. Um procedimento para avaliar as soluções, em cada geração, de acordo com a função objetivo do problema;

5. Escolha de parâmetros adequados para o tamanho da população, o número de gerações, e as respectivas probabilidades de cruzamento e mutação.

Estes componentes são descritos com mais detalhes na seção 5.6, que apresenta o desenvolvimento de um algoritmo genético combinado com o modelo Hipercubo

Alguns exemplos de aplicação bem sucedida de GA para solução de problemas de localização são os trabalhos de BEASLEY & CHU (1996), JARAMILHO et al. (2002), AYTUG & SAYDAM (2002), SAYDAM & AYTUG (2003) e COSTA (2004) .

BEASLEY & CHU (1996) desenvolveram um Algoritmo Genético para solucionar o problema de localização de cobertura de conjuntos. Neste estudo, os autores propõem um novo operador *crossover,* que eles chamam *crossover* por fusão, baseado no valor da função *fitness* dos cromossomos pais. Além disso, é também proposta uma taxa de mutação que varia de acordo com o comportamento de convergência do GA. Os resultados obtidos mostraram que o algoritmo é capaz de gerar soluções ótimas ou perto da ótima para os problemas testados.

Em JARAMILHO et al (2002), os autores aplicaram Algoritmos Genéticos para resolver um conjunto de problemas de localização. Neste estudo, os problemas escolhidos possuem métodos de solução conhecidos, cujo desempenho são bem documentados, o que proporcionou a comparação das soluções obtidas pelos mesmos com as soluções obtidas através da aplicação de GAs. O estudo mostrou que apenas para um problema analisado o desempenho do GA não é satisfatório. No entanto, para alguns problemas, o tempo computacional requerido pelo GA é significantemente superior ao requerido por outros métodos de solução.

AYTUG & SAYDAM (2002) compararam o desempenho da aplicação de GA para solucionar o MEXCLP com a heurística proposta por DASKIN (1983) e solução obtida por meio de um *software* de Programação Inteira. Os autores utilizaram a versão não linear da função objetivo para MEXCLP (NMEXCLP) proposta por SAYDAM & MACKNEW (1985). O estudo mostrou que o GA é capaz de produzir soluções ótimas ou perto da ótima em um tempo computacional razoável.

## 5.5 Abordagem combinando um algoritmo genético com o modelo Hipercubo (GA/Hipercubo)

Neste estudo desenvolvemos uma abordagem que integra o modelo Hipercubo com um algoritmo genético para determinar uma configuração ótima (ou perto da ótima), estabelecida pelo tamanho dos átomos do sistema. Os principais componentes considerados na implementação do algoritmo estão descritos nas subseções a seguir. Convém salientar que são muito poucos os trabalhos na literatura que estudam o melhor dimensionamento das áreas de preferenciais cada servidor sistemas emergenciais (*districting problem*). Além disso, desconhecemos estudos anteriores que tratam este problema nos SAEs em rodovias. Conforme mencionado neste capítulo, existem trabalhos na literatura combinando meta-heurísticas com o modelo Hipercubo para apoiar decisões de localização dos servidores em SAEs (p.e., SAYDAM & AYTUG (2003), CHIYOSHI et al. (2003) e GALVÃO et al. (2005)).

O primeiro passo da implementação do algoritmo GA/Hipercubo foi criar um procedimento que permita que novos dados de entrada sejam gerados para o modelo Hipercubo de acordo com diferentes configurações estabelecidas pela variação no tamanho de cada átomo do sistema. Para uma dada configuração, o algoritmo calcula as novas taxas de chegada de forma a preservar a distribuição original de demanda ao longo da rodovia. Uma nova matriz dos tempos de viagem entre os átomos do sistema também é calculada, dado que diferentes configurações apresentam diferentes posições de centróide para cada átomo. A nova taxa de chegada é calculada considerando cada par de átomos adjacentes da seguinte forma:

$$\text{Se } x_i < x_i^0 \text{ então } \begin{cases} \lambda_i = x_i\left(\lambda_i^0 / x_i^0\right), \\ \lambda_{i+1} = \lambda_{i+1}^0 + \left(x_{i+1} - x_{i+1}^0\right)\left(\lambda_i^0 / x_i^0\right) \end{cases}$$

$$\text{Se } x_i > x_i^0 \text{ então } \begin{cases} \lambda_i = \lambda_i^0 + \left(x_i - x_i^0\right)\left(\lambda_{i+1}^0 / x_{i+1}^0\right), \\ \lambda_{i+1} = x_{i+1}\left(\lambda_{i+1}^0 / x_{i+1}^0\right) \end{cases} \qquad (5.2)$$

onde $x_i^0$ é tamanho inicial do átomo *i*, $x_i$ é o tamanho final do átomo *i* (Figura 5.1), $\lambda_i^0$ é a taxa de chegada inicial do átomo *i*, $\lambda_i$ é taxa de chegada resultante do átomo *i* (*i* = 1, ...., $N_A$) e $d_j$ corresponde a distância entre dois servidores adjacentes $j$ e $j+1$, conforme ilustrado na figura 5.1.

### 5.5.1 Representação dos cromossomos:

A primeira representação de cromossomos, proposta por HOLLAND (1975), foi a representação binária 0-1. Entretanto, este tipo de representação não é adequado para todos os problemas combinatórios. Para certos problemas, a representação binária pode resultar em soluções infactíveis quando são aplicados os operadores genéticos (BEASLEY & CHU, 1996).

No presente estudo, cada gene do cromossomo é representado pela variável $y_j$, referente à proporção da distância entre dois servidores adjacentes (ou bases) *j* e *j+1* que corresponde a área primaria do servidor *j*. Lembre-se que, no sistema *Anjos do Asfalto*, exceto para os servidores das extremidades (1 e *N*), cada servidor possui dois átomos como área primária (os átomos da direita e esquerda). Desta forma, o número de genes em cada cromossomo é *N-1*, onde *N* é o número de servidores. A figura 5.1 ilustra a idéia geral desta representação, considerando as definições anteriores. Note que, $x_1 = y_1 d_1$ e $x_2 = d_1 - x_1$ para os átomos 1 e 2 entre as bases 1 e 2 adjacentes.

**Exemplo ilustrativo 7:**

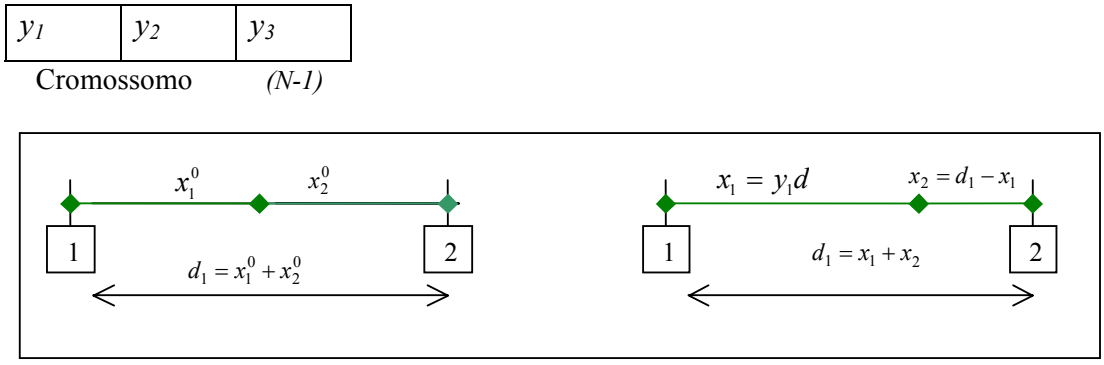


Figura 5.1 – Exemplo de variação do tamanho dos átomos entre dois servidores

Por esta representação, o gene $y_j$ gera a extensão dos átomos da área primária de cada servidor do sistema. Como mencionado anteriormente, diferentes configurações devem resultar em diferentes taxas de demanda nas áreas primárias e *backup* para cada servidor. E, conseqüentemente, diferentes medidas de desempenho devem ser obtidas para o sistema.

### 5.5.2 Procedimento de geração aleatória de cromossomos:

Dois algoritmos genéticos foram implementados, de forma a gerar os cromossomos de forma aleatória: um discreto e outro contínuo.

No caso discreto, para cada gene há um vetor de valores entre 0.2 e 0.8 que delimitam as partições da rodovia. Desta forma, a extensão de cada átomo primário de um dado servidor é limitada de 20 à 80 por cento da distância total aos seus servidores vizinhos. Esta delimitação (20/80 por cento) foi escolhida considerando que valores menores que 20 ou maiores que 80 são menos interessantes, pois resultariam praticamente, em eliminar uma área primária dos servidores adjacentes. O procedimento então determina como a distância entre dois servidores adjacentes deve ser particionada, partindo de 0.2 e adicionando um valor de incremento $\Delta$ (por exemplo, 0.01), multiplicado por um inteiro $k$, onde $k = 0 \ldots M = \left[\frac{0.6}{\Delta}\right]$. Portanto, para cada gene, atribui-se um número aleatório entre 0 e $M$ indexando o valor correspondente do vetor. Por exemplo, se o valor aleatório gerado para o gene 1 é 2 e $\Delta = 0.01$, então $k = 2$ e o valor do gene 1 é 0.22 (isto é, $0.2 + 2 \times 0.01$).

No caso contínuo, para cada gene é gerado um valor aleatório entre 0 e 1, e então multiplicado por 0.6. O valor obtido deve ser arredondado para a segunda casa decimal e finalmente somado a 0.2. Note que, como no caso anterior, o valor de cada gene varia entre 0.2 e 0.8. Por exemplo, se o valor aleatório na criação do gene 1 é aproximadamente 0.6324, temos que de acordo com os passos descritos acima, o valor atribuído ao gene 1 deve ser 0.58.

### 5.5.3 População Inicial:

A população inicial é gerada de forma aleatória de acordo com o procedimento descrito na seção 5.5.2. No primeiro algoritmo utilizamos o procedimento discreto para geração da primeira população e mutação, e no segundo algoritmo utilizamos o algoritmo contínuo.

### 5.5.4 Seleção dos cromossomos pais:

O principal objetivo da seleção em um Algoritmo Genético é prover a reprodução de soluções com alta avaliação (*fitness*). Os dois métodos mais utilizados na literatura de GA são: método da *roleta de probabilidades* e o método de *torneio.*

No método da *roleta de probabilidades*, os cromossomos pais são selecionados de acordo com a proporção de seu valor de aptidão (*fitness)*. O procedimento funciona como uma roleta, onde cada fatia representa a probabilidade de seleção de cada solução com base no valor de *fitness*. Como as melhores soluções apresentam fatias mais largas, ao rodar a roleta (executar a seleção), as melhores soluções têm maiores chances de serem selecionadas que as soluções com menor avaliação. O método de *torneio* corresponde à seleção aleatória de dois grupos de cromossomos com *T* membros. De cada grupo é escolhido um cromossomo pai com o melhor valor de aptidão para compor um par de pais no processo de *crossover* (cruzamento). Exemplos de aplicação bem sucedida deste método para problemas de localização são os trabalhos de BEASLEY & CHU (1996) e JARAMILHO et al. (2002).

No presente estudo, o método de *roleta de probabilidades* foi escolhido para as subseqüentes análises.

### 5.5.5 Avaliação e função *fitness*:

O procedimento de avaliação do presente algoritmo genético é baseado na inserção do modelo Hipercubo (descrito nos capítulos 3 e 4), para descrição das medidas de desempenho de interesse. Para cada cromossomo (solução), representando uma dada configuração do sistema, há uma taxa de chegada para cada átomo e uma matriz de tempos de viagem (de cada servidor a cada átomo) que correspondem aos dados de entrada do modelo Hipercubo. Através da aplicação deste modelo no procedimento de avaliação, obtemos as medidas de desempenho de cada solução individualmente na população.

Como mencionado anteriormente, o modelo Hipercubo fornece várias medidas interessantes que descrevem o desempenho do sistema. Entretanto, algumas destas medidas podem ser conflitantes em termos dos diferentes interesses envolvidos. Por exemplo, o balanceamento das cargas de trabalho dos servidores é uma medida de desempenho interna do sistema, que

interessa aos operadores do sistema. O gerente do sistema precisa otimizar a utilização dos recursos limitados. Por outro lado, o tempo médio de resposta na região é uma medida de desempenho externa do sistema, que interessa aos usuários do sistema.

Devido a este fato, inicialmente realizamos três experimentos com três diferentes funções de aptidão (*fitness)* que representam três objetivos: (i) tempo médio de viagem do sistema; (ii) fração de chamadas atendidas em mais de 10 minutos; (iii) desvio padrão das cargas de trabalho dos servidores.

No primeiro caso, o objetivo é oferecer um melhor nível de serviço ao usuário através da minimização do tempo médio de viagem no sistema. Como descrito pela equação (4.12) do capítulo 4, esta medida pode ser calculada e utilizada como função *fitness* da seguinte forma:

$$\overline{T}(c) = \sum_{j=1}^{N} \sum_{i=1}^{N_A} f_{ji} t_{ji} \ , \tag{5.3}$$

onde $\overline{T}(c)$ corresponde ao tempo médio de viagem no sistema para uma dada solução *c,* $f_{ji}$ é a fração de despachos do servidor $j$ ao átomo $i$ , obtida através da equação (4.10) do capítulo 4*,* e $t_{ji}$ é o tempo de viagem do servidor $j$ ao átomo *i.*

Para o segundo caso, a função *fitness* corresponde à fração de chamadas que são atendidas em tempo superior a 10 minutos. Uma das medidas de desempenho fornecidas pelo modelo Hipercubo é a fração de despachos de cada servidor a cada átomo. Se a distribuição dos tempos de viagem servidor – átomos for conhecida, podemos então calcular a proporção total de chamadas que requerem mais de 10 minutos para serem atendidas. O objetivo é minimizar o valor desta medida, definida por:

$$P_{tv>10}(c) = \sum_{j=1}^{N} \sum_{i=1}^{N_A} f_{ji}^{[v]} \ , \tag{5.4}$$

onde $P_{tv>10}(c)$ é a fração de chamadas que são atendidas em tempo superior à 10 minutos para a solução *c*, $f_{ji}^{[v]}$ é a fração de despachos do servidor *j* para o átomo *i*, que corresponde a um tempo de viagem superior à 10 minutos, e este termo é obtido da seguinte forma:

$$f_{ji}^{[v]} = f_{ji} p(t_{ji} > 10) \ \ , \tag{5.5}$$

onde $p(t_{ji} > 10)$ corresponde a probabilidade do tempo de viagem do servidor *j* ao átomo *i* ser maior que 10 minutos. A equação (5.6) é aplicada somente à sistemas que não admitem filas, onde $f_{ji} = f_{ji}^{[nq]}$ (calculado pela expressão (4.10) do capítulo 4).

No terceiro caso, o valor do desvio padrão entre as cargas de trabalho é utilizado para medir o desbalanceamento entre as mesmas, sendo que o objetivo é minimizar esta medida. Por meio do algoritmo genético combinado com o modelo Hipercubo, buscamos encontrar uma configuração que assegura um melhor balanço das cargas de trabalho. A função *fitness* é dada a seguir:

$$\sigma_\rho(c) = \sqrt{\frac{\sum(\rho_j(c) - \overline{\rho}(c))^2}{N}}, \qquad (5.6)$$

onde $\sigma_\rho(c)$ é o desvio padrão das cargas de trabalho da solução *c,* e $\rho_j(c)$ é calculado pela equação (4.7) do capítulo 4 e representa a carga de trabalho do servidor *j* na solução *c,* e $\overline{\rho}$ é a média das cargas de trabalho na solução *c*.

### 5.5.6 *Crossover*

O procedimento de *crossover* é fundamental na aplicação de GA e muitos pesquisadores defendem a idéia que este componente caracteriza um algoritmo como "genético" (GLOVER et al., 1995). Os métodos mais utilizados de aplicação deste operador são *crossover* de um ponto, o *crossover* de dois pontos e o *crossover* uniforme. O *crossover* de um ponto consiste na seleção aleatória de um gene (posição) no cromossomo e a criação de dois cromossomos filhos por meio da troca de genes entre dois cromossomos pais a partir deste ponto. O *crossover* de dois pontos funciona de forma similar, mas ao invés de um ponto, dois pontos são aleatoriamente escolhidos. No presente estudo, o *crossover* de um ponto é aplicado, como mostra a figura 5.3. Após a escolha aleatória de dois cromossomos pais, há uma probabilidade $p_c$ de que os mesmos irão passar pelo procedimento de cruzamento (*crossover*). Um número aleatório $r_c$ entre 0 e 1 é gerado e se $r_c \geq p_c$, o cruzamento ocorre, caso contrário os cromossomos pais são copiados para os cromossomos filhos, isto é, permanecem na população sem realizar o cruzamento.

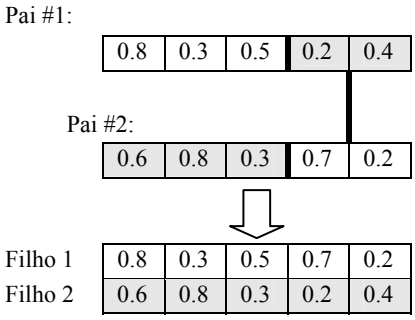


Figura 5.2 – *Crossover* de um ponto

### 5.5.7 Mutação:

A mutação é outro importante componente na geração de nova população a cada iteração da heurística GA. Na representação binária tradicional de um GA, a mutação é aplicada com uma pequena probabilidade através da inversão de um bit (1 para 0, ou 0 para 1) em um dado gene do cromossomo. A mutação é responsável pela diversificação dos cromossomos de forma a evitar que a busca chegue a um ponto de ótimo local.

Em nosso estudo, a mutação é aplicada a cada gene após o operador *crossover*, de acordo com uma probabilidade predefinida $p_m$. Assim, um número aleatório $r_m$ entre 0 e 1, é gerado e se $r_m \geq p_m$, então o gene é mutado. Para substituir os genes mutados, são escolhidos valores aleatórios por meio do procedimento descrito na seção 5.5.2. No algoritmo com população inicial gerada pelo método discreto, aplicamos também o método discreto para a mutação. No segundo algoritmo utilizamos o método contínuo para geração da população inicial e mutação.

### 5.5.8 Escolha dos parâmetros para aplicação do GA:

De acordo com a discussão anterior, a aplicação de um GA exige a escolha adequada da probabilidade de *crossover* ($p_c$), probabilidade de mutação ($p_m$), tamanho da população ($P_{op}$) e número de gerações ($G$). As respectivas probabilidades de *crossover* e mutação dependem do problema analisado. Para a presente análise, estes parâmetros foram escolhidos após testes extensivos com diferentes combinações de valores dentro de um intervalo. A melhor combinação resultante foi utilizada para a escolha dos demais parâmetros (tamanho da população e número de gerações). Os melhores resultados foram produzidos com a combinação: probabilidade de *crossover* ($p_c$) em torno de 0.7, probabilidade de mutação ($p_m$)

entre 0.05 e 0.07, tamanho da população $P_{op}$ = 100 e número de gerações $G$ = 1000 e $G$ = 2000.

Para o presente problema, um adicional parâmetro a ser avaliado corresponde ao intervalo Δ, para o caso de geração discreta de genes da população inicial e mutação. Os resultados obtidos variando este parâmetro dependem do tipo de função *fitness* testada.

### 5.5.9 Esquema básico do algoritmo GA/Hipercubo:

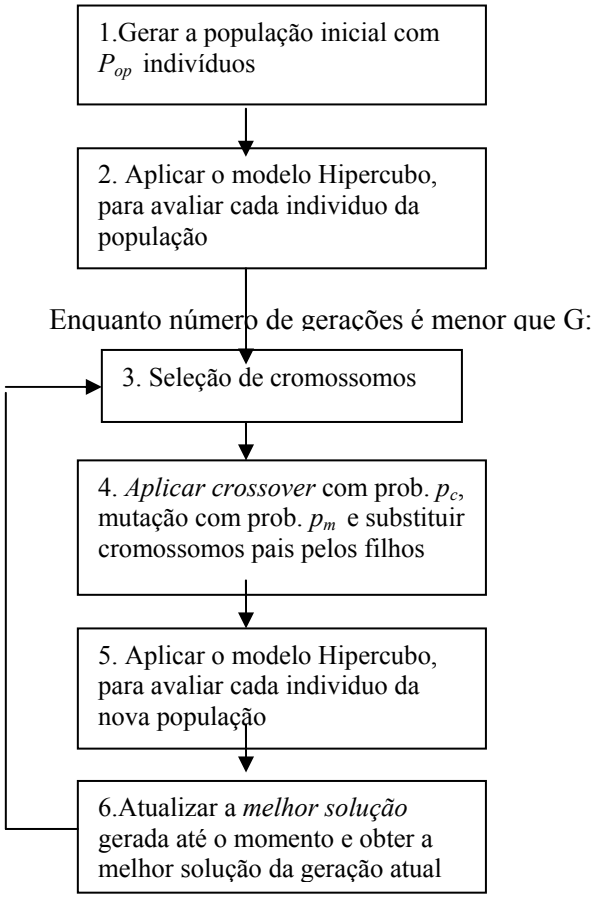


Figura 5.3 – Passos do Algoritmo GA/Hipercubo.

Passo 1: Uma população de *Pop* indivíduos (cromossomos) é gerada de forma aleatória, conforme descrito na seção 5.6.2;

Passo 2: Aplica-se o modelo Hipercubo a cada cromossomo (solução) *c* da população inicial, obtendo-se assim o valor da *fitness* associado ao mesmo;

Passo 3: Realiza-se a seleção dos pares de cromossomos que devem passar pelo procedimento de *crossover* (cruzamento) e/ou mutação para gerar a próxima população, conforme descrito na seção 5.6.3;

Passo 4: Os pares de cromossomos pais selecionados no passo 3 devem realizar o *crossover* com uma probabilidade $p_c$ . Se o *crossover* não ocorre, o par permanece na nova população e o tamanho da população é preservado. Os cromossomos resultantes do procedimento de *crossover*, isto é, cromossomos filhos e cromossomos pais (que não cruzaram) devem passar pelo procedimento de mutação, no qual cada um de seus genes são mutados de acordo com uma probabilidade $p_m$.

Passo 5: Após obtenção da nova população (geração), seus indivíduos são novamente avaliados. Conforme descrito na seção 5.6.4, no processo de avaliação, o modelo Hipercubo é aplicado a cada cromossomo *c*, e são obtidos o valor da *fitness* e outras medidas de desempenho.

Passo 6: Verifica-se se algum cromossomo da geração corrente tem o valor da *fitness* superior (melhor) que o da melhor solução até o momento encontrada. Neste caso, atualiza-se o valor da melhor solução. Verifica-se o critério de parada, estabelecido pelo número de gerações *G*. Se o número de gerações é menor que *G*, retorna-se ao passo 3.

Diversos autores, tais como JASZKIEWICZ (2002) e ARROYO (2002), têm sugerido incluir um procedimento de busca local para melhorar as soluções filhas geradas pelos operadores genéticos de *crossover* e mutação (algoritmo genético híbrido). Porém esta alternativa não foi considerada na presente abordagem, dado que o processo de avaliação de soluções baseado na solução do modelo Hipercubo, se tornaria computacionalmente "caro" para um dado número *G* de gerações necessárias.

**5.6 Otimização bi-objetivo** :

Como discutido anteriormente, as principais medidas de desempenho que queremos minimizar na configuração ótima do SAE analisado podem ser conflitantes quando tratamos o

problema com base em apenas um objetivo separadamente. Desta forma, o mais adequado seria tratar o problema como um problema multiobjetivo.

Como definido em ARROYO (2002), um problema de otimização multiobjetivo consiste em determinar um vetor de variáveis de decisão de forma a otimizar uma função vetorial composta por um conjunto de objetivos, que em geral são conflitantes, satisfazendo as restrições envolvidas. Por exemplo, um problema bi-objetivo com objetivos *f(x)* e *g(x)* pode ser formulado da seguinte forma:

$$\text{Min } Z = (f(x), g(x)) \qquad \text{(P1)}$$

$$\text{s.a } x \in X^*$$

onde $x$ corresponde ao vetor solução, $Z$ corresponde a imagem de $x$ ou espaço objetivo, e $X^*$ corresponde ao conjunto de soluções factíveis do problema.

Como não há uma solução ideal para o problema que otimize todos os objetivos, é necessário definir uma ordenação adequada das soluções factíveis do problema (ordenação parcial das soluções). Assim, dados dois vetores soluções $x$ e $y \in X^*$, estes podem ser relacionados de três formas com base nos seus vetores objetivo. Por exemplo, para o problema bi-objetivo (P1) acima temos :

(i) $f(x) < f(y)$ e $g(x) < g(y)$;
(ii) $f(x) > f(y)$ e $g(x) > g(y)$;
(iii) ou nenhum dos casos acima é verdadeiro;

Esta ordenação, introduzida pelo economista *Pareto* no século 19, determina a relação de dominância entre as soluções. Por exemplo, se $f(x) = 3$, $f(y) = 5$ e $g(x) = 4$ e $g(y) = 6$, as soluções $x$ e $y$ se relacionam em (i) e dizemos que $x$ domina $y$. Se $f(x) = 3$, $f(y) = 5$ e $g(x) = 4$ e $g(y) = 3$, $x$ e $y$ se relacionam em (iii) e são consideradas soluções indiferentes entre si, dado que uma não domina a outra (possuem o mesmo grau de dominância). Estas relações estão ilustradas no gráfico da figura 5.4 para o problema bi-objetivo (P1).

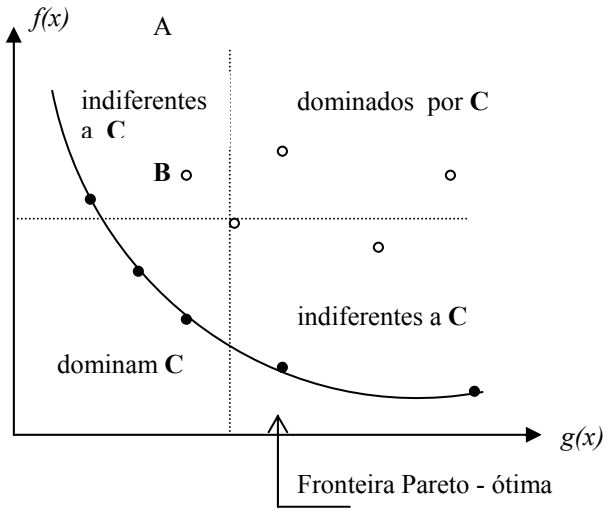


Figura 5.4 – Relação de dominância de Pareto

Na figura 5.4, para os pontos C e D, por exemplo, temos $f(C) < f(D)$ e $g(C) < g(D)$, então a solução C domina a solução D. No caso dos pontos B e C, temos $f(C) < f(B)$ e $g(B) < g(C)$, portanto as soluções B e C são indiferentes.

Uma solução $x^*$ é considerada eficiente ou Pareto-ótima, se não existe uma outra solução $x \in X^*$, que domine $x^*$. O conjunto de soluções eficientes do problema define a curva Pareto-ótima (linha entre os pontos pretos da figura 5.4).

De acordo com ARROYO (2002), os métodos de otimização multiobjetivo implicam em dois aspectos importantes: busca de soluções eficientes e tomada de decisões. Com base nestes aspectos, os métodos de otimização multiobjetivo são classificados em: (i) métodos *a-priori* (caracterizados pela participação dos decisores antes do processo de busca das soluções); (ii) métodos *a-posteriori* (o decisor seleciona uma solução adequada para o problema a partir de um conjunto de soluções eficientes) e (iii) métodos iterativos (o decisor atua durante o processo de busca de soluções, direcionando a para regiões onde existam soluções de interesse).

Os métodos mais tradicionais de otimização multiobjetivo, constam de transformar o problema multiobjetivo em um problema mono-objetivo, onde a otimização vetorial passa a ser escalar. Os métodos clássicos são apresentados, por exemplo, em COHON (1978) STEUER (1986) *apud* ARROYO (2002). Um dos mais simples métodos a priori de otimização multiobjetivo é o método da soma ponderada. Este consiste em combinar os

diferentes objetivos do problema em um mono-objetivo por meio de pesos de forma a estabelecer a preferência de cada objetivo. Por exemplo, o problema bi-objetivo (P1) é transformado em :

Minimizar $Z = w_1 f(x) + w_2 g(x)$ (P2)
s.a $x \in X^*$,
onde $w_i \geq 0$, e em geral $w_1 + w_2 = 1$

Este método é considerado um método *a- priori* dado que o decisor deve definir os pesos apropriados de acordo com a importância de objetivo.

Outra forma de resolver um problema *a priori*, é classificar os objetivos em ordem crescente de prioridade. E a partir desta classificação o problema é resolvido considerando apenas o primeiro objetivo, a seguir o problema é resolvido para o segundo objetivo utilizando como restrição o valor ótimo encontrado para o primeiro objetivo, e o processo continua até o último objetivo. Por exemplo, para o problema bi-objetivo (P1) temos:

Minimizar $Z = g(x)$
s.a $f(x) = f^*$, (P3)
$x \in X^*$, onde $f^*$ é a solução do problema:
Minimizar $Z = f(x)$
s.a $x \in X^*$

Como descrito em ARROYO (2002) e PILEGGI (2002), podemos também resolver um problema multiobjetivo por um método *a posteriori* que consta de minimizar (maximizar) um objetivo de maior prioridade sujeito a limitação dos outros objetivos (o método é chamado de método $\varepsilon$ –restrito). Por exemplo, para o problema bi-objetivo (P1) descrito acima, temos:

Minimizar $Z = f(x)$
s.a $g(x) \leq \varepsilon_g$ (P4)
$x \in X^*$ ou
Minimizar $Z = g(x)$
s.a $f(x) \leq \varepsilon_f$
$x \in X^*$

Ao variar de forma conveniente os limitantes $\varepsilon$, é possível gerar o conjunto Pareto-ótimo (conjunto de soluções eficientes). Um passo inicial importante é definir valores apropriados destes limitantes, de forma que o problema tenha solução.

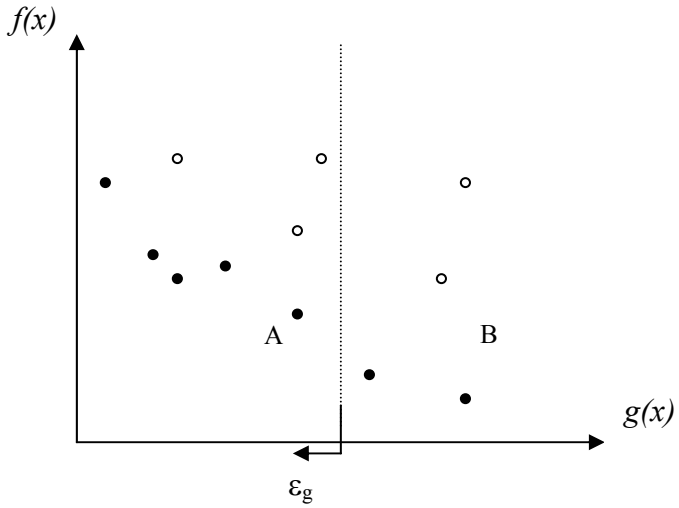


Figura 5.6 – Gráfico que ilustra o problema biobjetivo com restrição de um objetivo.

Note na figura 5.6, representando o espaço de soluções do problema (P1), que as soluções com $g(x) > \varepsilon_g$ não são consideradas soluções factíveis do problema, dado a restrição relacionada ao segundo objetivo. A solução A é solução ótima do problema (P4) e B é solução ótima do problema (P4) sem a restrição $g(x) \leq \varepsilon_g$.

Como descrito nas seções anteriores, o problema analisado no presente estudo deve ser solucionado como um problema multiobjetivo. Desta forma, podemos escolher duas medidas objetivo e tratar o problema como bi-objetivo, por exemplo, sendo o primeiro objetivo minimizar o desvio padrão das cargas de trabalho dos servidores $\sigma_\rho$, e o segundo minimizar o tempo médio de viagem no sistema $\overline{T}$, definidos conforme seção 5.6.4. Assim a função objetivo do problema bi-objetivo é: minimizar Z = $(f(x),g(x)) = (\sigma_\rho, \overline{T})$.

Neste estudo, utilizamos o método $\varepsilon$–restrito para solucionar o problema como mono-objetivo. Para isso devemos escolher um dos dois objetivos para ser minimizado sujeito a restrição do outro. Escolhemos arbitrariamente utilizar a medida de tempo médio de viagem $\overline{T}$ como restrição do problema, e otimizar o desvio padrão das cargas de trabalho dos servidores $\sigma_\rho$. Essa escolha é devida ao fato de que parece mais razoável para um tomador de decisão estabelecer limitantes $\varepsilon$ discretos para $\overline{T}$ (aqui denotado por $\overline{T}_{linit}$, por exemplo,

$\overline{T}_{linit}$=1,2,3,...minutos), do que fazer o mesmo para $\sigma_\rho$. Assim, o problema deve ser redefinido da forma:

Minimizar $Z = f(x) = \sigma_\rho$
s.a $g(x) = \overline{T} < \overline{T}_{linit}$
$x \in X^*$,
onde $\overline{T}_{linit}$ corresponde ao valor limitante superior de $\overline{T}$ estabelecido pelo usuário.

Uma vez definidos a função objetivo e a restrição do problema, utilizamos a abordagem GA/Hipercubo descrita na seção 5.4. para prescrever a solução aproximada do problema, sujeito ao valor de $\overline{T}_{linit}$.

# 6. Resultados

Neste capítulo apresentamos alguns resultados computacionais dos métodos propostos neste estudo. As seções 6.1, 6.2, 6.3 e 6.4 apresentam a aplicação do modelo Hipercubo para os dois estudos de caso de SAEs em rodovias com único e múltiplo despacho, respectivamente. Os resultados obtidos pelo modelo Hipercubo são também comparados com os resultados de modelos de simulação discreta. As seções 6.5 e 6.6 apresentam respectivamente os resultados obtidos pelo algoritmo genético combinado com o modelo Hipercubo exato, aplicado ao sistema *Anjos do Asfalto* e um procedimento iterativo para que integra o modelo Hipercubo para otimizar a configuração do sistema *Centrovias*.

Os algoritmos foram implementados em linguagem Pascal e executados em um microcomputador com processador Pentium de 2.0 GHz. Os modelos de simulação foram construídos e analisados utilizando o software *Arena* (SMC, 1994; KELTON et al. 2002) e executados em um microcomputador com processador Pentium de 345 MHz, que dispõe de licença para uso deste *software*. Os testes de aderência foram realizados pelo *software Best-Fit* (PC, 1996) e os testes de hipóteses foram realizados pelo *software Microcal Origin* (MICROCAL, 1997).

## 6.1 Modelo Hipercubo para análise do sistema *Anjos do Asfalto*

Nesta seção descrevemos os resultados da aplicação do modelo Hipercubo para análise do sistema *Anjos do Asfalto*, descrito na seção 2.2 do capítulo 2. As características de operação deste sistema são similares ao exemplo 3, pois este também é um SAE em rodovia com simples despacho (um servidor é despachado para cada chamada) e *backup* parcial. Portanto, a aplicação do modelo Hipercubo deve ser realizada conforme a discussão na seção 4.1 do capítulo 4.

Consideramos nesta aplicação, os dados de entrada coletados no sistema por MENDONÇA & MORABITO (2000, 2001), assim como a análise estatística apresentada naquele estudo. Estes dados são relacionados ao processo de chegada (intervalo entre chegadas das chamadas), ao processo de atendimento (tempo de atendimento das chamadas) e a localização dos

servidores. Como descrito pelos autores, os dados foram coletados em um período de pico de operação do sistema, que na época correspondia aos finais de semana.

No presente estudo analisamos também o tamanho de cada átomo, pois esta informação é utilizada na abordagem GA/Hipercubo proposta (veja seção 5.6 do capítulo 5). A partir da dos dados do tamanho de cada átomo e localização das bases de cada servidor, foi possível construir a matriz de tempo de viagem servidor – átomo. O tempo de viagem é calculado a partir da distância da base ao centróide do átomo, admitindo que a velocidade média dos veículos é 100 Km/h. Este valor parece ser razoável, dado que mesmo que o objetivo seja viajar na forma mais rápida possível, é preciso considerar que a velocidade deve ser reduzida devido ao peso dos veículos (ambulâncias são veículos relativamente pesados), e as condições de tráfego e relevo de certos trechos ao longo da viagem.

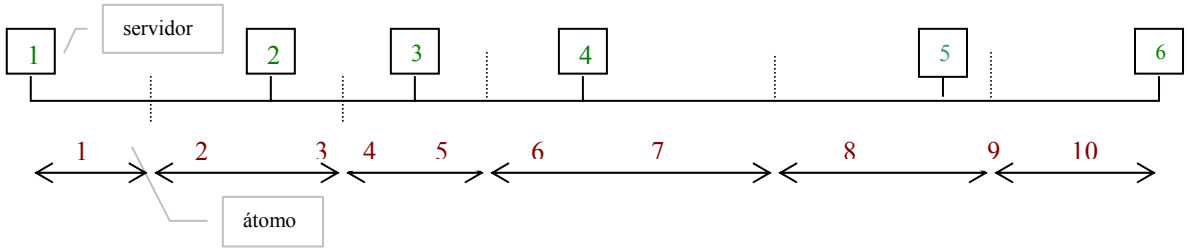


Figure 6.1 – Esquema da distribuição de átomos e servidores ao longo da rodovia no sistema Anjos do Asfalto

A figura 6.1 ilustra como os servidores e átomos estavam distribuídos ao longo do trecho da rodovia que correspondia ao SAE *Anjos do Asfalto*. A tabela 6.1 apresenta os dados de entrada para os 10 átomos do sistema. Nesta tabela temos: taxa média de chegada no sistema, lista de preferência (colunas 1 e 2), que corresponde aos 2 servidores preferenciais de cada átomo, e o tamanho (extensão em quilômetros) de cada átomo. A taxa total de chegadas no sistema é $\lambda = 0{,}01813$ chamadas/min (i.e, 1,0878 chamadas/hora).

Tabela 6.1 – Dados de entrada para cada átomo do sistema Anjos do Asfalto

| **Atomo** | **Taxa de chegada ($\lambda_i$) (chamadas/min)** | **Servidor preferencial** | **Servidor *backup*** | **Extensão (km)** |
|---|---|---|---|---|
| **1** | 0,00277 | 1 | 2 | 20.5 |
| **2** | 0,00084 | 2 | 1 | 20.5 |
| **3** | 0,00197 | 2 | 3 | 10.5 |
| **4** | 0,00111 | 3 | 2 | 10.5 |
| **5** | 0,00170 | 3 | 4 | 15.5 |
| **6** | 0,00008 | 4 | 3 | 15.5 |
| **7** | 0,00375 | 4 | 5 | 26.5 |
| **8** | 0,00184 | 5 | 4 | 26.5 |
| **9** | 0,00227 | 5 | 6 | 9.02 |
| **10** | 0,00180 | 6 | 5 | 31.98 |

Dado que, em MENDONÇA & MORABITO (2000, 2001) a taxa de chegada no átomo 6 foi suposta aproximadamente nula, adotamos um valor suficientemente pequeno igual a $\lambda_6 = 0.00008$. Desta forma, o átomo 6 adquire uma taxa de demanda suficientemente pequena, ou seja, não nula e menor que 10% da menor taxa de demanda dos átomos do sistema (átomo 2, $\lambda_2$=0,00084). Esta alteração foi necessária, para permitir a aplicação do algoritmo GA/Hipercubo como discutido na seção 5.6 do capítulo 5, dado que para $\lambda_6 = 0$, as variações do gene 2 (referentes a região entre os servidores 3 e 4) devem gerar diferentes configurações com mesmo valor de *fitness*. Para manter a taxa total de chegada igual a utilizada em MENDONÇA & MORABITO (2000, 2001) alteramos a taxa de chegada do átomo adjacente ao átomo 6 (átomo 5), de $\lambda_5 = 0{,}00178$ para $\lambda_5 = 0{,}00170$. Convém ressaltar que estas alterações são de segunda ordem e não alteram significantemente os resultados finais em relação aos resultados de MENDONÇA & MORABITO (2000, 2001).

A tabela 6.2 apresenta os dados de entrada relativos aos 6 servidores do sistema: taxa média de atendimentos por minuto e os átomos que são considerados área primária de cada servidor. A taxa total de atendimento é $\mu = 0{,}0959$, e portanto a taxa de ocupação do sistema é $\rho = \frac{\lambda}{\mu} = 0{,}1890$.

Tabela 6.2 – Dados de entrada dos servidores do sistema.

| **Servidor** | **Taxa de atendimento $\mu_j$ (chamadas/min)** | **Atomos da área primária** |
|---|---|---|
| **1** | 0,0187 | 1 |
| **2** | 0,0139 | 2, 3 |
| **3** | 0,0166 | 4, 5 |
| **4** | 0,0101 | 6, 7 |
| **5** | 0,0241 | 8, 9 |
| **6** | 0,0125 | 10 |

Note na tabela 6.2 que, com exceção dos átomos 1 e 2, cada servidor possui 2 átomos como área primária (ao lado esquerdo e direito de sua base).

A tabela 6.3 apresenta a matriz dos tempos médios de viagem entre a base do servidor $j$ e o átomo $i$: $t_{ji}$. Esta matriz é determinada no presente estudo a partir da distância (base do servidor - centróide do átomo), conforme discutido em LARSON & ODONI (1981). Analisando, por exemplo, o servidor 1 da figura 6.1, temos que o tempo de viagem deste servidor ao átomo 1 é calculado da seguinte forma:

Tamanho do átomo 1: $x_1$ = 20,5 km (tabela 6.1);

Distância do servidor 1 ao centróide do átomo 1 : $\frac{x_1}{2} = 10{,}25$ km;

Velocidade média = 100 km/h, então :

$t_{11} = 10{,}25 \cdot \left(\frac{60}{100}\right) = 6{,}15$ min (tabela 6.3);

O tempo de viagem do servidor 1 ao átomo 2 é obtido da seguinte forma:

$t_{12} = \left(x_1 + \frac{x_2}{2}\right) \cdot \left(\frac{60}{100}\right) = 18{,}45$ min (tabela 6.3).

Tabela 6.3 Tempo médio de viagem servidor – átomo (minutos): $t_{ji}$

| $t_{ji}$ (min) | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 6,150 | 18,450 | | | | | | | | |
| **2** | 18,450 | 6,150 | 3,150 | 9,450 | | | | | | |
| **3** | | | 9,450 | 3,150 | 4,650 | 13,950 | | | | |
| **4** | | | | | 13,950 | 4,650 | 7,950 | 23,850 | | |
| **5** | | | | | | | 23,850 | 7,950 | 2,706 | 15,006 |
| **6** | | | | | | | | | 21,894 | 9,594 |

**Equações de equilíbrio:**

Como o sistema *Anjos do Asfalto* possui N = 6 servidores, há $2^N = 64$ estados possíveis. As equações são determinadas de forma similar ao exemplo 3 do capítulo 4, considerando a lista de despacho da tabela 6.1. Por exemplo, para o estado $B = \{110001\}$, a equação de equilíbrio torna-se:

$$p_{110001}((\lambda_3 + \lambda_4 + \lambda_5 + \lambda_6 + \lambda_7 + \lambda_8 + \lambda_9 + \lambda_{10}) + \mu_1 + \mu_2 + \mu_6) = p_{100001}(\lambda_1 + \lambda_2 + \lambda_3) + p_{010001}(\lambda_1 + \lambda_2) + p_{110000}(\lambda_{10}) + p_{111001}(\mu_3) + p_{110101}(\mu_4) + p_{110011}(\mu_5) \quad (6.1)$$

Note que, o termo do lado esquerdo da equação (6.1), representando o fluxo para fora do estado B = {110001}, mostra que o sistema neste estado não pode atender chamadas geradas nos átomos 1 e 2, pois seus dois servidores preferencial e *backup* (1 e 2) estão ocupados.

No lado direito da equação (6.1), temos que as transições para dentro do estado B = {110001} são:

$\{100001\} \rightarrow \{110001\}$ – que ocorre com a chegada de uma chamada nos átomos 1( taxa $\lambda_1$ - atendimento *backup* ) , 2 ou 3 (taxa $\lambda_2 + \lambda_3$ - atendimento preferencial);

$\{010001\} \rightarrow \{110001\}$- que ocorre com a chegada de uma chamada nos átomos 1( taxa $\lambda_1$ - atendimento preferencial ) ou 2 (taxa $\lambda_2$ - atendimento *backup*);

$\{110000\} \rightarrow \{110001\}$- que ocorre com a chegada de uma chamada no átomo 10 ( taxa $\lambda_{10}$ - atendimento preferencial ), e

$\{111001\} \rightarrow \{110001\}$, $\{110101\} \rightarrow \{110001\}$, $\{110011\} \rightarrow \{110001\}$- que ocorrem com o término do serviço dos servidores 3, 4 e 5, respectivamente.

Com as probabilidades de equilíbrio computadas podemos calcular as principais medidas de desempenho deste sistema.

**Principais medidas de desempenho obtidas:**

Os resultados obtidos para as probabilidades de estados do sistema mostraram a probabilidade do estado $B = \{000000\}$ (i.e, todos os servidores disponíveis é $p_{\{00000\}} = 0{,}3085$ e do estado $B = \{111111\}$ (i.e, todos os servidores ocupados ) é $p_{\{11111\}} = 0{,}0001$.

Cargas de trabalho:

A tabela 6.4 apresenta os resultados obtidos para as cargas de trabalho de cada servidor $\rho_j$, obtidas de forma similar ao exemplo 3 do capítulo 4 (i.e, equação (4.7)), e também o desvio padrão que avalia o balanceamento das cargas de trabalho $\sigma_\rho$.

Tabela 6. 4 – Cargas de trabalho de cada servidor do sistema:

| **Servidor** $j$ | **Carga de trabalho** $\rho_j$ |
|---|---|
| **1** | 0,1352 |
| **2** | 0,1928 |
| **3** | 0,1612 |
| **4** | 0,3026 |
| **5** | 0,1833 |
| **6** | 0,1490 |
| $\sigma_\rho$ | 0,05507 |

Freqüências de despacho:

A tabela 6.5 apresenta os resultados para a freqüência de despacho de cada servidor a cada átomo: $f_{ji}$, calculada a partir da equação (4.10) do capítulo 4.

Tabela 6.5 Frequências de despacho servidor – átomo ( $f_{ji}$ )

| $f_{ji}$ | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 0,1391 | 0,0077 | | | | | | | | |
| **2** | 0,0161 | 0,0394 | 0,0924 | 0,0078 | | | | | | |
| **3** | | | 0,0174 | 0,0541 | 0,0828 | 0,0012 | | | | |
| **4** | | | | | 0,0106 | 0,0032 | 0,1519 | 0,0117 | | |
| **5** | | | | | | | 0,0499 | 0,0873 | 0,1077 | 0,0117 |
| **6** | | | | | | | | | 0,0192 | 0,0890 |

Note que, $\sum_{j}^{6}\sum_{i}^{10} f_{ji} = 1$.

Probabilidade de Perda:

A probabilidade de perda do sistema ( $P_p$ ) é igual à 0,05, sendo obtida de forma similar à expressão (4.8) para o exemplo 3 do capítulo 4. A fração de chamadas atendidas em mais que 10 minutos ( $P_{tr>10}$ ), calculada pela expressão (5.11) do capítulo 5, é 12,81%.

Tempo de viagem:

O tempo médio de viagem no sistema $\overline{T}$, calculado por meio da expressão (4.12) do capítulo 4, é 7,9121 minutos. A tabela 6.6 apresenta os tempos médios de viagem para cada átomo $i$ do sistema ($\overline{T_i}$), calculados por meio da expressão (4.13).

Tabela 6.6 – Tempo médio de viagem para cada átomo (minutos).

| **Atomo** | **Tempo médio de viagem $\overline{T_i}$ (min)** |
|---|---|
| **1** | 7,4258 |
| **2** | 8,1597 |
| **3** | 4,1481 |
| **4** | 3,9410 |
| **5** | 5,7066 |
| **6** | 7,0958 |
| **7** | 11,8824 |
| **8** | 9,8352 |
| **9** | 5,6121 |
| **10** | 10,2210 |

A tabela 6.7 apresenta os tempos meio de viagem de cada servidor $j$ do sistema ($\overline{TU_j}$), obtidos por meio da expressão (4.14) do capítulo 4.

Tabela 6.7 – Tempo médio de viagem para cada servidor (minutos)

| **Servidor** | **Tempo de viagem (min)** |
|---|---|
| **1** | 6,7943 |
| **2** | 5,8067 |
| **3** | 4,7343 |
| **4** | 9,3003 |
| **5** | 9,1631 |
| **6** | 11,779 |

Os resultados de $\overline{T_i}$ e $\overline{TU_j}$ foram comparados com a amostra e validados em MENDONÇA & MORABITO (2000,2001), e os desvios obtidos foram da ordem de 7%. Como admitimos algumas modificações com relação aos dados originais (p.e, a taxa de chegada no átomo 6 e calculo da matriz de tempo de viagem servidor - átomo), procuramos também validar o modelo através de simulação.

Os resultados para as medidas de desempenho do sistema *Anjos do Asfalto* são utilizados na seção 6.3, que descreve os resultados da implementação do algoritmo GA/Hipercubo para este sistema.

**Modelo de Simulação:**

Um modelo de simulação foi construído para o sistema *Anjos do Asfalto*, permitindo verificar se as simplificações admitidas na implementação do modelo Hipercubo não influem na acuracidade da análise deste sistema. Um modelo de simulação do sistema *Anjos do Asfalto* foi construído utilizando os dados de entrada apresentados nas tabelas 6.1 à 6.3.

A simulação considera chegadas Poisson independentes em cada átomo, e a mesma política de despacho discutida na seção 2.3, do capítulo 2. No entanto, foi necessário considerar uma aproximação para atribuir a distribuição dos tempos de atendimento de cada servidor. Lembrando que, uma das hipóteses do modelo Hipercubo é que o tempo de atendimento $1/\mu_j$ inclui o tempo de *set-up* (preparação), o tempo de viagem (ida e volta) e o tempo em cena, no modelo de simulação, calculamos de forma separada o tempo de viagem de ida (do local do servidor ao local do chamado), e o tempo de atendimento para o servidor *j* é dado por:

$$Ts_j = T_{0j} - t_{ji}, \qquad (6.2)$$

onde $Ts_j$ é o tempo de atendimento do servidor *j* utilizado como dado de entrada do modelo de simulação, $T_{0j} = 1/\mu_j$ é o tempo de atendimento dos dados coletados do sistema e utilizado como dado de entrada no modelo Hipercubo (tabela 6.2) e $t_{ji}$ é o tempo de viagem do servidor *j* ao átomo *i* (tabela 6.3).

Cálculo da fase transiente e do tempo total de simulação:

A fase transiente (*warm up*) foi determinada através da ferramenta *Output Analyzer* do software *Arena*. Para medida a ser analisada como resultado da simulação, utilizamos a análise do gráfico da média móvel dos resultados obtidos em uma longa rodada de simulação para determinar período transiente. Para determinar o tempo total de simulação, utilizamos o Método de Loteamento (LAW & KELTON, 1991; PEDGEN et al, 1995, KELTON et al., 2002). Por este método, determinamos o tamanho dos lotes de observações e o número de lotes necessários para garantir uma correlação próxima de zero entre estes, similarmente ao

procedimento relatado em IANNONI & MORABITO (2002, 2004). Este procedimento deve ser realizado para cada resultado (medida) da simulação individualmente, pois cada medida possui um tempo entre observações. Por exemplo, há grandes diferenças entre o número de observações para o tempo médio de viagem para chamadas entre um átomo com alta taxa de chegada e um átomo com relativamente pequena taxa de chegada. Além disso, o número de observações é um dado importante para o cálculo do intervalo de confiança e da correlação no Método de Loteamento. Assim, o tempo de simulação foi determinado para cada medida da seguinte forma: (i) determinar qual o número de observações necessárias em cada lote que garanta correlação próxima de zero; (ii) determinar qual o intervalo de tempo de simulação entre observações; e (iii) multiplicar este intervalo pelo número de lotes, o número de observações em cada lote e o fator de segurança (PEDGEN et al, 1995), obtendo o tempo total de simulação sem a fase transiente.

Podemos também escolher uma medida que possua maior correlação entre os lotes ou que possua o maior intervalo entre chamadas, realizar o procedimento de loteamnento descrito acima com base nesta medida, e verificar se o número de observações para as demais medidas é de fato suficiente para determinação do intervalo de confiança e garantia de correlação próxima de zero. Utilizando o software Arena, está verificação pode ser feita diretamente pelo relatório de resultados que reporta as informações de intervalo de confiança, insuficiência de observações e correlação.

No caso do sistema *Anjos do Asfalto* há medidas cujas observações ocorrem dentro de intervalos muito longos. Observe por exemplo na tabela 6.1, que a taxa de chamada no átomo 6 é $\lambda_6 = 0.00008$ e portanto uma observação neste átomo ocorre a cada 12.500 minutos. Escolhendo esta medida para aplicação dos procedimentos descritos acima, destes procedimentos, observamos que o período transiente está em torno de 10.000 minutos. O tempo total de simulação do modelo de simulação, incluindo a fase transiente é de: 3.218.000 minutos. Note que, o tempo total de simulação é relativamente alto.

As tabelas 6.8 a 6.10 comparam os resultados obtidos pelo modelo Hipercubo e pelo modelo de simulação para tempo médio de viagem para cada átomo do sistema, cargas de trabalho dos servidores e tempo médio de viagem para cada servidor. A quarta coluna compara estes resultados por meio do desvio relativo entre os valores médios de cada medida. A quinta coluna reporta o intervalo de confiança (COSTA,1977; KELTON et al., 2002) para os

resultados da simulação, calculado com nível de significância $\alpha$ = 95%. Nesta análise verificamos se os resultados obtidos pelo Hipercubo estão dentro do intervalo de confiança dos resultados da simulação. .

Tabela 6.8 – Tempo médio de viagem para cada átomo (minutos).

| **Átomo** $i$ | $\overline{T_i}$ **(min) – modelo** | $\overline{T_i}$ **(min) - simulação** | **Intervalo de confiança (simulação)** |
|---|---|---|---|
| **1** | 7,426 | 7,422 | 7,342 – 7,502 |
| **2** | 8,160 | 8,145 | 7,989 – 8,300 |
| **3** | 4,1481 | 4,151 | 4,086 – 4,217 |
| **4** | 3,9410 | 3,913 | 3,231 – 4,594 |
| **5** | 5,707 | 5,686 | 5,670 – 5,702 |
| **6** | 7,096 | 6,986 | 6,568 – 7,404 |
| **7** | 11,882 | 11,858 | 11,737 – 11,978 |
| **8** | 9,835 | 9,790 | 9,668 – 9,121 |
| **9** | 5,612 | 5,661 | 5,474 – 5,848 |
| **10** | 10,221 | 10,227 | 10,267 – 10,187 |

Tabela 6.9 – Resultados para cargas de trabalho (modelo x simulação):

| **Servidor** $j$ | $\rho_j$ **- modelo** | $\rho_j$ **- simulação** | **Intervalo de confiança (simulação)** |
|---|---|---|---|
| **1** | 0,1352 | 0,1337 | 0,1309 – 0,1365 |
| **2** | 0,1928 | 0,1927 | 0,1880 – 0,1974 |
| **3** | 0,1612 | 0,1602 | 0,1567 – 0,1637 |
| **4** | 0,3026 | 0,3048 | 0,2980 – 0,3116 |
| **5** | 0,1833 | 0,1830 | 0,1794 – 0,1866 |
| **6** | 0,1490 | 0,1492 | 0,1459 – 0,1525 |

Tabela 6.10 – Tempo médio de viagem para cada servidor (minutos)

| **Servidor** $j$ | $\overline{TU_j}$ **(min) – modelo** | $\overline{TU_j}$ **(min) – simulação** | **Intervalo de confiança (simulação)** |
|---|---|---|---|
| **1** | 6,794 | 6,803 | 6,745 – 6,860 |
| **2** | 5,807 | 5,782 | 5,669 – 5,894 |
| **3** | 4,734 | 4,741 | 4,696 – 4,786 |
| **4** | 9,300 | 9,250 | 9,171 – 9,328 |
| **5** | 9,163 | 9,155 | 9,011 – 9,328 |
| **6** | 11,779 | 11,700 | 11,614 – 11,786 |

Note que os desvios relativos apresentados nas tabelas 6.8 a 6.10 entre resultados do modelo para os valores médios dos resultados obtidos na simulação são pouco significativos. O desvio médio para o tempo médio de viagem para cada átomo é de 1,13%, o desvio médio para cargas de trabalho dos servidores é de 1,36%, e para o tempo médio de viagem dos servidores é de 0,53%.

Com relação a comparação por meio do intervalo de confiança, note nas tabelas 6.8 a 6.10 que algumas medidas apresentam intervalo de confiança relativamente pequeno, pois o número de observações para estas medidas na longa rodada de simulação é relativamente grande (muito maior que outras medidas). Ao analisar o intervalo de confiança para cada medida obtida pela simulação, verificamos que sempre os resultados obtidos pelo Hipercubo estão dentro deste intervalo. Portanto, o modelo pode ser considerado validado pelo modelo de simulação.

## 6.2 Modelo Hipercubo para análise do SAE *Centrovias*

### Coleta de Dados

A pesquisa de campo no SAE da *Centrovias* foi realizada em duas etapas: na primeira etapa foram realizadas visitas gerais ao SAE, à C.C.O e a sede da *Centrovias* em São Carlos. Nestas primeiras visitas foram coletadas informações sobre o funcionamento do sistema como: número de servidores, disposição dos servidores, características gerais dos SAUs, política de despacho, entre outras informações fundamentais.

A segunda etapa foi realizada com diversas visitas à C.C.O, localizada em Itirapina, para coleta de dados dos relatórios preenchidos pelos resgatistas dos SAUs e arquivados pelos operadores da C.C.O. As principais informações destes relatórios são: número do evento, data do evento, localização, instante de acionamento, instante de chegada no local, instante de chegada no hospital (quando ocorre), instante de chegada de volta a base, tipo de evento, descrição do evento e do atendimento.

Os dados coletados correspondem ao período de dezembro 2001 à junho 2002. Durante este período não foram constatados períodos de pico ou significantes variações ao longo dos meses analisados. Com relação aos dias da semana, o gráfico da figura 6.2 apresenta a distribuição dos eventos em cada dia da semana durante todo o período de dezembro a junho.

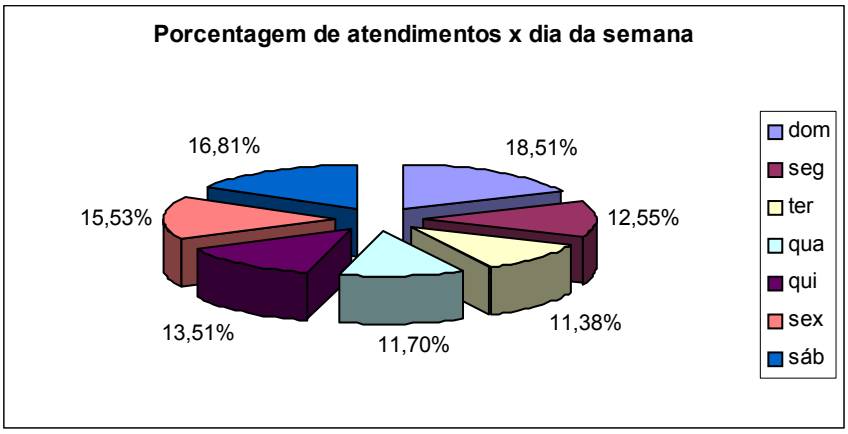


Figura 6.2 : Distribuição de eventos ao longo da semana

Note na figura 6.2 que há poucas diferenças entre os dias da semana, porém as maiores proporções ocorrem no final de semana (sexta à domingo). Este resultado também foi observado em MENDONÇA & MORABITO (2000) no estudo dos *Anjos do Asfalto*. No presente estudo, isto pode ser explicado pelo aumento do fluxo de veículos nos finais de semana, entre a capital (São Paulo) e o interior. Realizamos um teste de hipóteses (COSTA NETO, 1977) para verificar se a taxa média de atendimento nos finais de semana é igual a taxa média de atendimento durante a semana. O teste com nível de significância $\alpha = 0{,}05\%$, não rejeitou a hipótese de taxas médias iguais.

Os dados foram coletados e analisados considerando também o funcionamento contínuo do sistema durante 24 horas do dia e os 7 dias da semana. Ao todo foram contados 945 eventos durante este período.

**Divisão dos trechos de rodovia em átomos:**

A divisão do trecho em átomos foi realizada de acordo com as informações fornecidas pelos operadores a respeito da política de despacho adotada e da área de atendimento preferencial de cada servidor (veja seção 2.4 do capítulo 2). Para análise deste sistema, inicialmente dividimos o trecho em 8 átomos, como apresentado na figura 6.3, que se diferenciam em tamanho, lista de preferência de despacho e taxa de chegada. O tamanho e a lista de preferência de despacho de cada átomo (servidores preferencial e *backup*) estão na tabela 6.11.

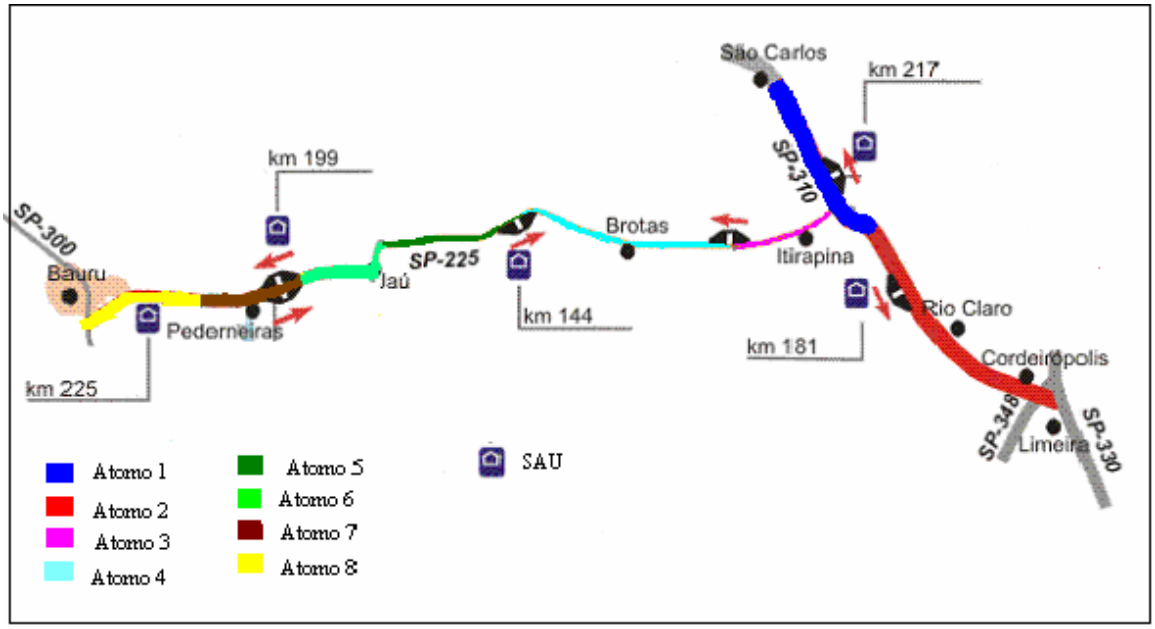


Figura 6.3 :Divisão do trecho em átomos (Fonte: Mapa fornecido pela Centrovias, e aqui modificado c/ a divisão de átomos)

Tabela 6.11 Extensão e lista de preferência de cada átomo (*backup* parcial)

| **Átomo** $i$ | **Extensão (Km)** | **Servidor preferencial** | **Servidor *backup*** |
|---|---|---|---|
| **1** | 32 | 1 | 2 |
| **2** | 42 | 2 | 1 |
| **3** | 15 | 1 | 2 |
| **4** | 38 | 3 | 1 |
| **5** | 26 | 3 | 2 |
| **6** | 29 | 4 | 3 |
| **7** | 16 | 4 | 5 |
| **8** | 20 | 5 | 4 |

Note que, como mencionado no capítulo 2, no SAE *Centrovias* cada átomo possui 2 servidores em sua lista de despacho (servidor preferencial e servidor *backup*). No entanto, há servidores que podem viajar para atender uma chamada em até 4 átomos (servidores 1 e 2), enquanto que há servidor que viaja para somente para dois átomos (servidor 5).

**Análise Estatística dos dados:**

Processo de Chegada

A tabela 6.12 apresenta os resultados obtidos da análise do processo de chegada de chamadas em cada átomo do sistema. Estes dados são: número de observações em cada átomo, intervalo médio entre chegadas, desvio padrão, coeficiente de variação (razão entre o desvio padrão e a

média), taxa de chegada $\lambda_i$ e fração de múltiplo despacho do total de atendimentos em cada átomo ($pr_i^{[2]}$). Assim, a taxa de chegada de chamadas tipo 2 em cada átomo $i$ é $\lambda_i^{[2]} = \lambda_i \cdot pr_i^{[2]}$ e a taxa de chegada de chamadas tipo 1 no átomo $i$ é $\lambda_i^{[1]} = \lambda_i - \lambda_i^{[2]}$.

Tabela 6.12 – Dados do processo de chegada

| **Átomo** $i$ | **Número de observações** | **Intervalo médio (horas)** | **Desvio padrão (horas)** | **Coef. de variação** | **Taxa de chegada** $\lambda_i$ | **Fração de múltiplo despacho** |
|---|---|---|---|---|---|---|
| **1** | 326 | 15,336 | 17,094 | 1,115 | 0,0652 | 0,0521 |
| **2** | 267 | 18,887 | 18,674 | 0,988 | 0,0529 | 0,0899 |
| **3** | 37 | 136,525 | 108,978 | 0,798 | 0,0073 | 0,0540 |
| **4** | 49 | 96,325 | 68,839 | 0,715 | 0,0104 | 0,0408 |
| **5** | 28 | 202,464 | 255,579 | 1,262 | 0,0049 | - |
| **6** | 55 | 87,793 | 71,081 | 0,810 | 0,0114 | - |
| **7** | 74 | 69,435 | 63,340 | 0,912 | 0,0144 | - |
| **8** | 109 | 46,889 | 49,987 | 1,066 | 0,0213 | 0,0917 |

A taxa total de chegadas no sistema é $\lambda = 0{,}1878$ chamadas/hora. Observe que nos dados analisados não foram encontrados eventos de múltiplo despacho nos átomos 5, 6 e 7, portanto, $\lambda_5^{[2]} = \lambda_6^{[2]} = \lambda_7^{[2]} = 0$. Os coeficientes de variação são razoavelmente próximos de 1.0, o que sugere que a distribuição dos intervalos entre chegadas possa ser aproximadamente exponencial.

A figura 6.4 abaixo mostra a proporção de chamadas em cada em átomo do total de chamadas atendidas no sistema:

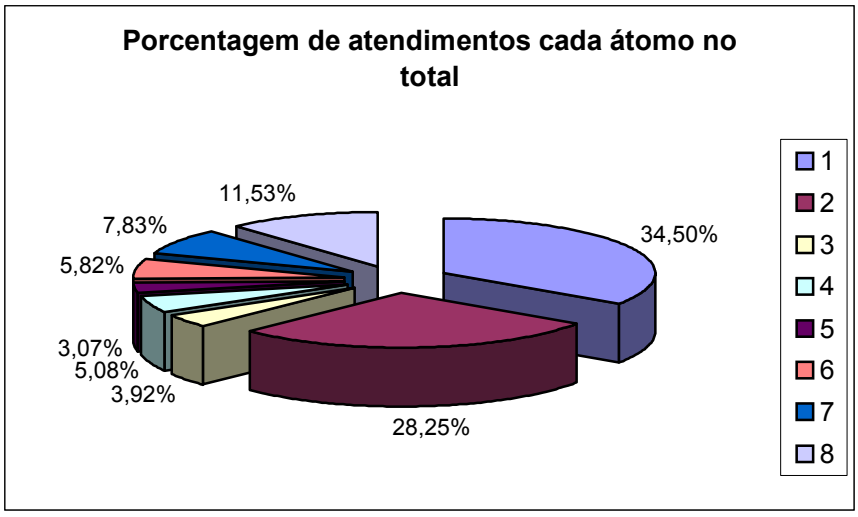


Figura 6.4 – Proporção de eventos em cada átomo

Note nos dados da tabela 6.12 e figura 6.4 que a maior proporção de atendimentos ocorre nos átomos 1 e 2, que correspondem a trechos da rodovia Washington Luis, enquanto que a menor taxa de chamadas ocorre nos átomos 3 e 4.

Teste de aderência: Foram realizadas análises estatísticas para verificar se o processo de chegada é Poisson. Os testes utilizados pelo software *Best-Fit* são: *Kolmogorov-Smirnov, Anderson Darling, Qui-quadrado*. Os resultados obtidos desta análise, com nível de significância $\alpha = 0{,}05$, mostraram que, para todos os átomos, não se pode rejeitar a hipótese de que os intervalos entre chegadas sucessivas tem distribuição exponencial. Este resultado também foi obtido nas análises de MENDONÇA & MORABITO (2000, 2001) para o processo de chegada do sistema *Anjos do Asfalto*.

Portanto, neste trabalho considerou-se que o número de chegadas por intervalo de tempo tem distribuição de Poisson.

Processo de Atendimento

A tabela 6.13, a seguir, apresenta os resultados obtidos da análise do processo de atendimento realizado por cada servidor do sistema. Estes dados são: tempo médio de atendimento em minutos, desvio padrão, o coeficiente de variação (razão entre o desvio padrão e a média), taxa média de atendimento e fração de ocupação (obtida através da razão entre a soma total do tempo de atendimento do servidor e o tempo total de operação). O valor de $\mu$ é 5,4321. Portanto, $\rho = \frac{\lambda}{\mu} = 0{,}0346$.

Tabela 6.13 – Dados do processo de atendimento

| **Átomo** $i$ | **Número de observações** | **Tempo atendimento (horas)** | **Desvio padrão (minutos)** | **Coef. de variação** | **Taxa de atendimento** $\mu_j$ | **Fração de ocupação** |
|---|---|---|---|---|---|---|
| **1** | 390 | 0,7642 | 0,5308 | 0,6946 | 1,3085 | 0,0587 |
| **2** | 287 | 0,9397 | 0,5179 | 0,5511 | 1,0642 | 0,0532 |
| **3** | 75 | 1,2126 | 0,7788 | 0,6422 | 0,8247 | 0,0182 |
| **4** | 140 | 0,9215 | 0,3584 | 0,3889 | 1,0852 | 0,0252 |
| **5** | 108 | 0,8699 | 0,6194 | 0,7120 | 1,1495 | 0,0185 |

Note que os coeficientes de variação (coluna 5 da tabela 6.13) não são tão próximos de 1.0, como observado na tabela 6.12. Observe na última coluna da tabela 6.13, que os servidores tem pequena taxa de ocupação. Note também que, os servidores 1 e 2 (localizados na rodovia

Washington Luis) possuem maior carga de trabalho, dado que suas áreas de atendimento correspondem ás áreas com maior taxa de chegadas (veja também a tabela 6.12 e figura 6.4).

O tempo de atendimento compreende o tempo de *set-up* , o tempo viagem da base (SAU) ao local do evento, o tempo em cena, o tempo de viagem ao hospital, o tempo de volta à base. O tempo de *set-up* neste sistema é em média de 1 minuto, como regra de operação para os resgatistas.

Para verificar a variabilidade nos tempos de atendimento entre os servidores, realizamos a análise de variância ANOVA (COSTA NETO, 1977; MAGALHÃES & LIMA, 2002), com nível de significância $\alpha = 0{,}05$. O resultado desta análise mostrou que as diferenças entre as médias dos tempos de atendimento entre os servidores são significantes. Portanto, a aplicação do modelo Hipercubo deve considerar que os servidores não são homogêneos (i.e, as taxas de atendimento $\mu_j$ são distintas).

Teste de aderência: Como no processo de chegada, também foram realizadas análises estatísticas para verificar se a distribuição dos tempos de atendimento pode ser considerada exponencial. Os resultados obtidos nesta análise, mostraram que esta hipótese deve ser rejeitada, com nível de significância $\alpha = 0{,}05$, similarmente ao sistema *Anjos do Asfalto* (MENDONÇA & MORABITO, 2000, 2001). Porém, como comentado na seção 3.3 do capítulo 3 e em JARVIS (1985), este tipo de sistema pode ser analisado aproximadamente pelo modelo Hipercubo sem que a análise seja comprometida.

Matriz dos Tempos de viagem:

Para o cálculo dos tempos de viagem, utilizamos as informações sobre a velocidade média de cada servidor ($\bar{v}_j$). Onde $\bar{v}_1 = 95{,}08$, $\bar{v}_2 = 81{,}64$, $\bar{v}_3 = 100{,}80$, $\bar{v}_4 = 80{,}64$ e $\bar{v}_5 = 78{,}20$ km/h. Utilizando a velocidade média de cada servidor e a distância entre o local da chamada e base de cada servidor (SAU), foi possível obter o tempo médio de viagem de cada servidor a cada átomo (no caso átomos em que o servidor pode atender) considerando que há dois tipos de despacho: chamadas tipo 1 (único despacho) e chamadas tipo 2 (duplo despacho). Desta forma, a tabela 6.14 mostra a matriz do tempo médio de viagem servidor – átomo para chamadas tipo 1 ($t_{ji}^{[1]}$) e a tabela 6.15 apresenta esta matriz para chamadas tipo 2 ($t_{ji}^{[2]}$).

Note que, para as chamadas tipo 1, substituímos $t_{ji}$ nas equações (4.37) e (4.38), do capítulo 4 por $t_{ji}^{[1]}$. Para chamadas tipo 2, substituímos $t_{ji}$ nas equações (4.39) à (4.46), do capítulo 4 por $t_{ji}^{[2]}$. Este tratamento é aqui necessário, pois os tempos médios de cada servidor a cada átomo para chamadas tipo 1 e tipo 2 apresentam significante variabilidade para alguns servidores. Note, por exemplo, nas tabelas 6.14 e 6.15 que $t_{11}^{[1]} = 4{,}76$ e $t_{11}^{[2]} = 10{,}06$ (mais que o dobro que $t_{11}^{[1]}$). Um dos principais fatores que levam a estas diferenças é que significante proporção das chamadas tipo 1 são atendidas na própria base (i.e, $t_{ji} = 0$), como mostra a tabela 6.16. A influência da proporção de chamadas atendidas na base é relevante pelo fato que, diferentemente do sistema *Anjos do Asfalto*, o cálculo do tempo de viagem não utiliza a distância entre as bases e o centroíde de cada átomo, mas sim o local exato de cada chamada dentro do átomo.

Um outro fator que pode contribuir para que $t_{ji}^{[2]} > t_{ji}^{[1]}$, é o fato de que, para certos acidentes, uma ambulância *backup* pode reduzir a velocidade se souber que há outra se dirigindo ao local do acidente.

Tabela 6.14 – Tempo de viagem servidor – átomo $t_{ji}^{[1]}$ (minutos)

| $t_{ji}^{[1]}$ | **Átomo 1** | **Átomo 2** | **Átomo 3** | **Átomo 4** | **Átomo 5** | **Átomo 6** | **Átomo 7** | **Átomo 8** |
|---|---|---|---|---|---|---|---|---|
| SAU 1 | 4,76 | 15,99 | 13,21 | 22,30 | - | - | - | - |
| SAU 2 | 16,18 | 6,63 | *23,18 | - | - | - | - | - |
| SAU 3 | - | - | - | 6,65 | 5,75 | 24,11 | - | - |
| SAU 4 | - | - | - | - | **40,92 | 7,05 | 5,92 | **19,35 |
| SAU 5 | - | - | - | - | - | - | *13,81 | 3,50 |

* calculados com base na distância do servidor ao centróide do átomo;
** apenas uma observação.

Na tabela 6.14, os valores de $t_{23}^{[1]}$ e $t_{57}^{[1]}$ foram calculados através da distância base do servidor ao centróide do átomo, dado que não ocorreu despachos tipo 1 do servidor 2 ao átomo 3, e do servidor 5 ao átomo 7. O valor de $t_{45}^{[1]}$ ocorre para apenas um despacho observado entre servidor 4 e o átomo 5.

Tabela 6.15 – Tempo de viagem servidor – átomo para tipo 2 $t_{ji}^{[2]}$ (minutos)

| $t_{ji}^{[2]}$ | **Átomo 1** | **Átomo 2** | **Átomo 3** | **Átomo 4** | **Átomo 5** | **Átomo 6** | **Átomo 7** | **Átomo 8** |
|---|---|---|---|---|---|---|---|---|
| SAU 1 | 10,06 | 20,51 | *7,89 | *21,46 | - | - | - | - |
| SAU 2 | 13,76 | 6,36 | *16,94 | - | - | - | - | - |
| SAU 3 | - | - | - | *18,45 | - | - | - | - |
| SAU 4 | - | - | - | - | - | - | - | 13,84 |
| SAU 5 | - | - | - | - | - | - | - | 4,63 |

* apenas duas observações.

Note na tabela 6.15 que os átomos 5, 6 e 7 não apresentam atendimentos múltiplo despacho observados, e os valores de $t_{13}^{[2]}$ , $t_{14}^{[2]}$, $t_{23}^{[2]}$ e $t_{34}^{[2]}$ são obtidos de apenas dois eventos observados em cada caso. Pode-se observar nas tabelas 6.13, 6.14 e 6.15, que os tempos de viagem correspondem à pequena proporção do tempo de atendimento.

A tabela 6.16 a seguir mostra a localização (km na rodovia) e proporção de chamadas atendidas na base (tempo de viagem nulo) de cada SAU.

Tabela 6.16 –Localização e fração de atendimento na base do servidor (SAU)

| **Servidor** $j$ | **Localização** | **Fração de atendimento na base** |
|---|---|---|
| SAU 1 | Km 217-SP310 | 22,99% |
| SAU 2 | Km 187-SP310 | 11,46% |
| SAU 3 | Km 144-SP225 | 17,10% |
| SAU 4 | Km 199-SP225 | 17,73% |
| SAU 5 | Km 224-SP225 | 2,78% |

Note que a taxa de atendimentos na base é significante, principalmente para o SAU 1 (mais de 20%).

**Aplicação do Modelo Hipercubo :**

O modelo Hipercubo de múltiplo despacho para SAEs em rodovias, que foi discutido capítulo 4, foi implementado para analisar o estudo de caso da concessionária *Centrovias*. Os resultados preliminares obtidos para as principais medidas de desempenho são apresentados a seguir.

Equações de Equilíbrio:

As equações de equilíbrio são determinadas de forma similar ao exemplo ilustrativo 4 do capítulo 4, que representa um SAE em rodovias com política de múltiplo despacho. Dado que o SAE *Centrovias* tem 5 servidores, então há $2^5 = 32$ estados possíveis do sistema.

Analisando por exemplo, o estado B = {11001}, com base na lista de preferência de despacho da tabela 6.11, temos a seguinte equação de equilíbrio:

$$\begin{aligned} &p_{11001}((\lambda_4^{[1]} + \lambda_5^{[1]} + \lambda_6^{[1]} + \lambda_7^{[1]} + \lambda_8^{[1]} + \lambda_4^{[2]} + \lambda_5^{[2]} + \lambda_6^{[2]} + \lambda_7^{[2]} + \lambda_8^{[2]}) + \mu_1 + \mu_2 + \mu_5) = \\ &p_{00001}(\lambda_1^{[2]} + \lambda_2^{[2]} + \lambda_3^{[2]}) + p_{10001}(\lambda_1^{[1]} + \lambda_2^{[1]} + \lambda_3^{[1]} + \lambda_1^{[2]} + \lambda_2^{[2]} + \lambda_3^{[2]}) + \\ &p_{01001}(\lambda_1^{[1]} + \lambda_2^{[1]} + \lambda_3^{[1]} + \lambda_1^{[2]} + \lambda_2^{[2]} + \lambda_3^{[2]}) + p_{11000}(\lambda_8^{[1]}) + p_{11101}(\mu_3) + p_{11011}(\mu_4) \end{aligned} \qquad (6.3)$$

Observe que, de acordo com o lado esquerdo da equação (6.3), o sistema não permite que chamadas tipo 1 ou tipo 2 dos átomos 1, 2 e 3 sejam atendidas, pois neste estado os 2 servidores preferenciais destes átomos estão ocupados. Embora o servidor 5 esteja ocupado, chamadas de ambos os tipos nos átomos 7 e 8 são atendidas pelo servidor 4.

Com relação ao lado direito da equação (6.3), temos que as transições para dentro do estado B = {11001} são:

{00001} $\rightarrow$ {11001} – que ocorre com a chegada de uma chamada tipo 2 nos átomos 1, 2 ou 3 (taxa total $\lambda_1^{[2]} + \lambda_2^{[2]} + \lambda_3^{[2]}$);

{10001} $\rightarrow$ {11001}- que ocorre com a chegada de uma chamada tipo 1 ou tipo 2 nos átomos 1, 2 ou 3 (taxa total $\lambda_1^{[1]} + \lambda_2^{[1]} + \lambda_3^{[1]} + \lambda_1^{[2]} + \lambda_2^{[2]} + \lambda_3^{[2]}$);

{01001} $\rightarrow$ {11001}- que ocorre com a chegada de uma chamada tipo 1 ou tipo 2 nos átomos 1, 2 ou 3 (taxa total $\lambda_1^{[1]} + \lambda_2^{[1]} + \lambda_3^{[1]} + \lambda_1^{[2]} + \lambda_2^{[2]} + \lambda_3^{[2]}$);

{11000} $\rightarrow$ {11001}- que ocorre com a chegada de uma chamada tipo 1 no átomo 8 (taxa $\lambda_8^{[1]}$);

{11101} $\rightarrow$ {11001}, {11011} $\rightarrow$ {11001}- que ocorrem com o término do serviço dos servidores 3 e 4, respectivamente.

**Resultados para medidas de desempenho obtidas pelo modelo:**

Após determinar as probabilidades de equilíbrio, é possível determinar as medidas de desempenho do sistema relacionadas aos atendimentos com único e múltiplo despacho, tais como: carga de trabalho, fração de despachos à chamadas tipo 1 e tipo 2, probabilidades de perda e tempos de viagem. Esta análise é similar a realizada para o exemplo 4 do capítulo 4 e é apresentada a seguir.

Os resultados obtidos para as probabilidades de estados do sistema mostraram que o mesmo tem probabilidade muito pequena de estar saturado (todos servidores ocupados), pois $p_{\{11111\}}$ = 0,00000034. Na maior parte do tempo o sistema permanece ocioso (todos servidores livres), pois $p_{\{00000\}}$ = 0,8434.

Carga de Trabalho dos SAUs

As cargas de trabalho de cada servidor $\rho_j$ foram calculadas de forma similar ao exemplo 4 do capítulo 4 (i.e, equação (4.21)), e os resultados são apresentados na tabela 6.17. Na coluna 3 desta tabela estão a fração de ocupação obtida da amostra de dados, e a coluna 4 apresenta o desvio entre o resultado do modelo e da amostra.

Tabela 6.17 – Resultados para carga de trabalho dos servidores

| **Servidor** | $\rho_j$ **Modelo** | $\rho_j$ **Amostra** | **Desvio relativo** |
|---|---|---|---|
| SAU 1 | 0,0578 | 0,0587 | -1,53% |
| SAU 2 | 0,0537 | 0,0532 | 0,94% |
| SAU 3 | 0,0186 | 0,0182 | 2,19% |
| SAU 4 | 0,0253 | 0,0252 | -0,39% |
| SAU 5 | 0,0185 | 0,0185 | 0% |

Note que os desvios dos resultados do modelo para a amostra são pequenos, e o desvio médio é de apenas 1,0%.

Probabilidade de Perda:

A probabilidade de perda para chamadas tipo 1 ($P_p^{[1]}$) é igual à 0,00590 e para chamadas tipo 2 ($P_p^{[2]}$) é igual à 0,00680. A probabilidade de perda para o sistema ($P_p$) é 0,00595. Estas

medidas foram calculadas de forma similar ao exemplo 4 do capítulo 4 (i.e, expressões (4.22) à (4.24), respectivamente).

Freqüências de despacho:

Como descrito no capítulo 3 e 4, as freqüências de despacho para um sistema com múltiplo despacho podem ser diferenciadas, por exemplo, em: freqüências de despacho para chamadas tipo 1 (simples despacho), freqüências de despacho chamadas tipo 2 que são atendidas por 2 servidores, freqüências de despacho para chamada tipo 2 que são atendidas por um servidor e freqüências de despacho total (tipo 1 e tipo 2) de cada servidor a cada átomo.

(i) Freqüências de despacho tipo 1:

A tabela 6.18, a seguir, apresenta as freqüências de despacho do tipo 1 de cada servidor a cada átomo considerando o total de despachos tipo 1 no sistema, ou seja $f_{ji}^{[1]}$. Estes resultados são obtidos pela expressão (4.25) do capítulo 4.

Tabela 6.18 – Freqüência de despacho tipo 1 servidor $j$ – átomo $i$ ($f_{ji}^{[1]}$)

| $f_{ji}^{[1]}$ | **Átomo 1** | **Átomo 2** | **Átomo 3** | **Átomo 4** | **Átomo 5** | **Átomo 6** | **Átomo 7** | **Átomo 8** |
|---|---|---|---|---|---|---|---|---|
| SAU 1 | 0,3310 | 0,0124 | 0,0371 | 0,0010 | - | - | - | - |
| SAU 2 | 0,0174 | 0,2592 | 0,0019 | - | - | - | - | - |
| SAU 3 | - | - | - | 0,0556 | 0,0276 | 0,0016 | - | - |
| SAU 4 | - | - | - | - | 0,0005 | 0,0631 | 0,0798 | 0,0019 |
| SAU 5 | - | - | - | - | - | - | 0,0019 | 0,1081 |

Verifique que, se somarmos as freqüências da tabela 6.15, temos que $\sum_{j}^{5}\sum_{i}^{8} f_{ji}^{[1]} = 1$. Note também que, embora nos dados coletados não foi observado ocorrência de despacho tipo 1 do servidor 2 ao átomo 3, o modelo estima que $f_{23}^{[1]} = 0{,}0019$.

(ii) Freqüências de despacho tipo 2:

As freqüências de chamadas tipo 2 com despacho de dois servidores $f_{(j,k)i}^{[2]}$ para cada átomo $i$, onde o servidor $j$ é o servidor preferencial e o servidor $k$ é o servidor *backup* deste átomo, foram obtidas pela expressão (4.28) do capítulo 4. Lembre-se que nessa expressão $f_{(j,k)i}^{[2]}$ é

calculada considerando o total de despachos tipo 2 no sistema. Assim temos: $f^{[2]}_{(1,2)1} = 0{,}2806$, $f^{[2]}_{(2,1)2} = 0{,}3929$, $f^{[2]}_{(1,2)3} = 0{,}0330$, $f^{[2]}_{(3,1)4} = 0{,}0358$, $f^{[2]}_{(3,4)5} = 0$, $f^{[2]}_{(4,3)6} = 0$, $f^{[2]}_{(4,5)7} = 0$ e $f^{[2]}_{(5,4)8} = 0{,}1728$.

Os demais valores de $f^{[2]}_{(j,k)i}$ são nulos, pois somente os 2 servidores preferenciais de cada átomo podem atendê-lo. Note que, $f^{[2]}_{(3,4)5} = 0$, $f^{[2]}_{(4,3)6} = 0$ e $f^{[2]}_{(4,5)7} = 0$, pois a taxa de chamadas tipo 2 nos átomos 5, 6 e 7 é nula (tabela 6.12).

As freqüências de despacho tipo 2 com despacho de um único servidor $f^{[2]}_{ji}$, obtidos pela expressão (4.30) são apresentados na tabela 6.19.

Tabela 6.19 – Freqüência de despacho tipo 2 - despacho de 1 servidor $j$ – átomo $i$ ($f^{[2]}_{ji}$)

| $f^{[2]}_{ji}$ | **Átomo 1** | **Átomo 2** | **Átomo 3** | **Átomo 4** | **Átomo 5** | **Átomo 6** | **Átomo 7** | **Átomo 8** |
|---|---|---|---|---|---|---|---|---|
| SAU 1 | 0,0142 | 0,0199 | 0,0017 | 0,0007 | - | - | - | - |
| SAU 2 | 0,0155 | 0,0217 | 0,0018 | - | - | - | - | - |
| SAU 3 | - | - | - | 0,0022 | - | - | - | - |
| SAU 4 | - | - | - | - | - | - | - | 0,0031 |
| SAU 5 | - | - | - | - | - | - | - | 0,0043 |

A soma das freqüências $f^{[2]}_{ji}$ da tabela 6.19 e os valores de $f^{[2]}_{(j,k)i}$ é dada por: $\sum_{i=1}^{8}\left[\sum_{j=1}^{5} f^{[2]}_{ji} + \sum_{j=1}^{4}\sum_{k=j+1}^{5} f^{[2]}_{(j,k)i}\right] = 1$, de acordo com a expressão (4.32) do capítulo 4.

Como discutido no capítulo 3, podemos também obter freqüências de despachos com base no total de despachos (chamadas tipo 1 e chamadas tipo 2). Assim, os valores de $f'^{[1]}_{ji}$ correspondem à freqüência de todos os despachos do sistema que envia o servidor $j$ ao átomo $i$, para atender uma chamada do tipo 1, e foram calculados pela expressão (4.33) do capítulo 4. Os resultados são apresentados na tabela 6.20.

Tabela 6.20 – Freqüência de despacho tipo 1 servidor $j$ – átomo $i$ ( $f'^{[1]}_{ji}$ )

| $f'^{[1]}_{ji}$ | **Átomo 1** | **Átomo 2** | **Átomo 3** | **Átomo 4** | **Átomo 5** | **Átomo 6** | **Átomo 7** | **Átomo 8** |
|---|---|---|---|---|---|---|---|---|
| SAU 1 | 0,3118 | 0,0117 | 0,0350 | 0,0009 | - | - | - | - |
| SAU 2 | 0,0164 | 0,2441 | 0,0018 | - | - | - | - | - |
| SAU 3 | - | - | - | 0,0523 | 0,0260 | 0,0015 | - | - |
| SAU 4 | - | - | - | - | 0,0005 | 0,0594 | 0,0751 | 0,0018 |
| SAU 5 | - | - | - | - | - | - | 0,0018 | 0,1018 |

Os valores de $f'^{[2]}_{(j,k)i}$ correspondem à freqüência de todos os despachos do sistema que envia o servidor $j$ e $k$ ao átomo $i$ para uma chamada tipo 2, onde o servidor $j$ é o servidor preferencial e o servidor $k$ é o servidor *backup* deste átomo, obtidos pela expressão (4.34) do capítulo 4 são: $f'^{[2]}_{(1,2)1} = 0{,}0163$, $f'^{[2]}_{(2,1)2} = 0{,}0019$, $f'^{[2]}_{(1,2)3} = 0{,}0229$, $f'^{[2]}_{(3,1)4} = 0{,}0021$, $f'^{[2]}_{(3,4)5} = 0$, $f'^{[2]}_{(4,3)6} = 0$, $f'^{[2]}_{(4,5)7} = 0$ e $f'^{[2]}_{(5,4)8} = 0{,}0100$. Como mencionado anteriormente, os demais valores de $f'^{[2]}_{(j,k)i}$ são nulos, pois somente os 2 servidores preferenciais de cada átomo podem atendê-lo.

A tabela 6.21 apresenta os valores de $f'^{[2]}_{ji}$ calculados pela expressão (4.35) , que correspondem à freqüência de todos os despachos do sistema que envia o servidor $j$ ao átomo $i$, para atender uma chamada do tipo 2.

Tabela 6.21 – Freqüência de despacho tipo 2 - despacho de 1 servidor $j$ – átomo $i$, com base em todos os despachos ( $f'^{[2]}_{ji}$ )

| $f'^{[2]}_{ji}$ | **Átomo 1** | **Átomo 2** | **Átomo 3** | **Átomo 4** | **Átomo 5** | **Átomo 6** | **Átomo 7** | **Átomo 8** |
|---|---|---|---|---|---|---|---|---|
| SAU 1 | 0,0008 | 0,0012 | 0,0001 | 0,00004 | - | - | - | - |
| SAU 2 | 0,0009 | 0,0013 | 0,0001 | - | - | - | - | - |
| SAU 3 | - | - | - | 0,0001 | - | - | - | - |
| SAU 4 | - | - | - | - | - | - | - | 0,0002 |
| SAU 5 | - | - | - | - | - | - | - | 0,0002 |

Note que temos, conforme expressão (3.46) do capítulo 3:

$$\sum_{i=1}^{8}\left[\sum_{j=1}^{N}(f'^{[1]}_{ji}+f'^{[2]}_{ji})+\sum_{j=1}^{4}\sum_{k=j+1}^{5}f'^{[2]}_{(j,k)i}\right]=1\,.$$

Tempo de viagem:

(i) Atendimentos tipo 1 (único despacho):

Tempo médio de viagem no sistema :

O tempo médio de viagem no sistema para chamadas tipo 1 ($\overline{T}^{[1]}$) deve ser calculado pela expressão (4.37) do capítulo 4, substituindo $t_{ji}$ por $t_{ji}^{[1]}$. O resultado obtido pelo modelo é $\overline{T}^{[1]}$ = 6,277 minutos. O resultado obtido pela análise da amostra de dados é 5,962 minutos. Portanto temos um desvio de apenas 5,3% do resultado do modelo para o resultado da amostra (tabela 6.22).

Tempo médio de viagem para cada servidor:

O tempo médio de viagem para cada servidor considerando apenas despacho tipo 1 ($\overline{TU}_j^{[1]}$) é calculado pela expressão (4.38) do capítulo 4, substituindo $t_{ji}$ por $t_{ji}^{[1]}$. Os valores desta medida são apresentados na tabela 6.22.

Tabela 6.22 – Tempo de médio viagem para cada servidor – único despacho $\overline{TU}_j^{[1]}$ (minutos)

| **Servidor** $j$ | $\overline{TU}_j^{[1]}$ **Modelo** | $\overline{TU}_j^{[1]}$ **Amostra** | **Desvio** |
|---|---|---|---|
| SAU 1 | 5,993 | 5,667 | 5,75% |
| SAU 2 | 7,342 | 6,748 | 8,80% |
| SAU 3 | 6,686 | 6,652 | 0,51% |
| SAU 4 | 6,705 | 6,748 | 0,64% |
| SAU 5 | 3,682 | 3,508 | 4,96% |
| $\overline{T}^{[1]}$ | 6,277 | 5,962 | 5,30% |

Note na tabela 6.22 que os desvios dos resultados do modelo são relativamente pequenos com relação aos valores obtidos pela análise da amostra de dados. O desvio relativo médio para os resultados de $\overline{TU}_j^{[1]}$ é de 3,8%.

(ii) Atendimentos tipo 2:

Tempo médio de viagem no sistema para chamadas do tipo 2 (considerando o primeiro veículo a chegar no sistema) - $\overline{T}^{[2]}$

O tempo médio de viagem no sistema para chamadas tipo 2 ($\overline{T}^{[2]}$) é obtido pela expressão (4.39) do capítulo 4, substituindo $t_{ji}$ por $t_{ji}^{[2]}$. O valor obtido pelo modelo para este medida é 8,186 minutos.

Na análise dos dados da amostra não foi possível identificar quais são os atendimentos tipo 2 realizados por apenas um servidor, ou seja, $f_{ji}^{[2]}$ e $f'^{[2]}_{ji}$. Por isso o valor de $\overline{T}^{[2]}$, que no modelo considera os atendimentos tipo 2 realizados por um e dois servidores, não pôde ser diretamente comparado com o valor obtido da amostra, calculado considerando apenas atendimentos tipo 2 que são atendidos por 2 servidores. Assim, para comparar $\overline{T}^{[2]}$ obtido pelo modelo, com o valor obtido pela amostra, redefinimos $\overline{T}^{[2]}$ da expressão (4.39) para:

$$\overline{T}'^{[2]} = \frac{\sum_{i=1}^{N_A}\sum_{j=1}^{N-1}\sum_{k=j+1}^{N} f_{(j,k)i}^{[2]} \min(t_{ji}, t_{ki})}{\sum_{i=1}^{N_A}\sum_{j=1}^{N-1}\sum_{k=j+1}^{N} f_{(j,k)i}^{[2]}} \tag{6.4}$$

O resultado obtido é 7,776 minutos, e o valor da amostra é 7,201 minutos (desvio relativo de 7,9 %).

Tempo médio de viagem no sistema para chamadas do tipo 2 (incluindo todos os servidores despachados ) - $\overline{T}_t^{[2]}$

Utilizando a expressão (4.40) do capítulo 4 para calcular $\overline{T}_t^{[2]}$ pelo modelo, obtemos 24,067 minutos. De forma similar ao caso anterior, devemos redefinir $\overline{T}_t^{[2]}$ para realizar a comparação com o resultado da amostra. A expressão (4.40) torna-se então:

$$\overline{T}_t'^{[2]} = \frac{\sum_{i=1}^{N_A}\sum_{j=1}^{N-1}\sum_{k=j+1}^{N} f_{(j,k)i}^{[2]} (t_{ji} + t_{ki})}{\sum_{i=1}^{N_A}\sum_{j=1}^{N-1}\sum_{k=j+1}^{N} f_{(j,k)i}^{[2]}} \tag{6.5}$$

O resultado obtido no modelo pela expressão (6.5) é 25,130 minutos, e o valor obtido pela análise da amostra é 25,144 minutos (i.e, o resultado do modelo apresenta um desvio de apenas 0,056% com relação à amostra).

Como pode ser observado, o desvio de $\overline{T}'^{[2]}$ para a amostra (7,9%) é significantemente maior que o de $\overline{T}_t'^{[2]}$ (0,056%). Isso se deve ao fato de que, pela amostra podemos identificar e computar no cálculo de $\overline{T}^{[2]}$ em quais atendimentos o servidor preferencial foi o segundo a chegar. Enquanto que no modelo, utilizamos as médias dos tempos de viagem de cada

servidor a cada átomo (tabela 6.15), e portanto, o servidor preferencial é sempre o primeiro a chegar no local do evento. Tais simplificações influenciam o cálculo de $\overline{T}^{[2]}$ e os desvios obtidos, o que não ocorre com $\overline{T}_t^{[2]}$. Note que na expressão (6.5), o tempo de viagem dos dois veículos (primeiro e segundo a chegar) são considerados.

Tempo médio de viagem para o servidor preferencial das chamadas tipo 2:

Esta medida é calculada pela expressão (4.41) do capítulo 4, que considera os despachos em que o servidor $j$ é o servidor preferencial quando é acompanhado pelo servidor $k$, e os despachos em que o servidor $j$ é despachado como servidor preferencial, pois é o único disponível dos 2 servidores que podem atender o átomo $i$ ( $j$ pode ser o 1º ou o 2º servidor da lista de preferência de $i$ ). O valor desta medida obtida pelo modelo é 7,474 minutos.

De forma similar às duas medidas anteriores, é necessário modificar a expressão (4.41) para comparar o resultado obtido pelo modelo com a análise da amostra. Assim a expressão (4.41) torna-se:

$$\frac{\sum_{i=1}^{N_A}\sum_{j=1}^{N}\sum_{k\in L_i'} f_{(j,k)i}^{[2]} t_{ji}}{\left(1-\sum_{j=1}^{N}\sum_{i=1}^{N_A} f_{ji}^{[2]}\right)} \qquad (6.6)$$

Note que de acordo com a discussão anterior, no modelo, a expressão (6.6) equivale à expressão (6.4), pois pela tabela (6.15), observamos que o servidor preferencial é sempre o primeiro a chegar no local do evento. Portanto o valor da expressão (6.6) é 7,776 minutos, e o resultado obtido pela análise da amostra é 7,771 minutos. O desvio obtido é de apenas 0,064%, pois como no cálculo de $\overline{T}_t^{[2]}$, o fato de que o servidor preferencial nem sempre é o primeiro a chegar no local do evento não influencia o cálculo desta medida.

Tempo médio de viagem para o servidor *backup* das chamadas tipo 2 :

O valor desta medida, calculado no modelo pela expressão (4.42), é 15,880 minutos. Para comparar o modelo com a amostra recalculamos este valor através da expressão (4.42) modificada para:

$$\frac{\sum_{i=1}^{N_A}\sum_{j=1}^{N}\sum_{k\in L_i} f_{(j,k)i}^{[2]} t_{ki}}{\left(1-\sum_{j=1}^{N}\sum_{i=1}^{N_A} f_{ji}^{[2]}\right)} \qquad (6.7)$$

O valor obtido pela expressão (6.7) é 17,373 minutos e pela amostra é 17,447 minutos (desvio de apenas 0,4%).

O tempo médio de viagem para o primeiro servidor a chegar no local de uma chamada tipo 2 ($\overline{T}_F$) calculado pela expressão (4.43), apresenta os mesmos resultados da expressão (6.4), pois consideramos apenas chamadas tipo 2 , que são atendidas por dois servidores. O tempo médio de viagem para o segundo servidor a chegar no local de uma chamada tipo 2 ($\overline{T}_S$), calculado no modelo pela expressão (4.44), apresenta os mesmos resultados da expressão (6.7). O valor obtido pela amostra é 17,944 minutos, sendo que o desvio do modelo para a amostra é de 3,3%. Note que neste caso o desvio também depende do fato que, na amostra há eventos em que o servidor *backup* é o primeiro a chegar no local do acidente.

Fração de chamadas atendidas em mais que 10 minutos:

Utilizando a expressão (4.47) do capítulo 4 , calculamos a fração de chamadas atendidas em tempo superior a 10 minutos: $P_{tv>10} = \sum_{j=1}^{N}\sum_{i=1}^{N_A} f_{ji}^{v}$ . Esta medida foi calculada utilizando os dados da amostra da distribuição de chamadas atendidas em tempo maior que 10 minutos para cada servidor em cada átomo. Os resultados obtidos pelo modelo mostraram que para 26,2% de todas as chamadas, 25,7% das chamadas tipo 1 e 33,5% das chamadas tipo 2, o tempo de resposta é maior que 10 minutos.

**Modelo de Simulação:**

Os resultados do modelo Hipercubo também são comparados com os resultados obtidos por um modelo de simulação do SAE *Centrovias*. O modelo foi construído de forma similar ao modelo de simulação para o sistema *Anjos do Asfalto*. Porém, como a pesquisa de campo deste estudo de caso foi realizada totalmente neste estudo, podemos obter através da análise dos dados coletados o tempo de atendimento para cada servidor sem o tempo de viagem.

Assim, o tempo de atendimento médio de cada servidor é obtido considerando: o tempo de *set-up*, o tempo de atendimento em cena, o tempo de viagem ao hospital (se necessário) e o tempo de volta à base. Além disso, neste modelo consideramos que em cada átomo o intervalo entre chamadas tem distribuição exponencial e há dois tipos de chamadas (tipo 1 e tipo 2), conforme tabela 6.12.

O período transiente (*warm-up*) e o tempo total de simulação foram determinados pelo mesmo método descrito na seção 6.1 para o sistema *Anjos do Asfalto*. Dado que em alguns átomos o intervalo médio entre chegadas é longo, o tempo total da rodada de simulação para estas medidas deve também ser suficientemente longo. Isso porque, deve ser gerado um número suficiente de observações para todas as medidas analisadas, de forma obter um intervalo de confiança e garantir que correlação entre os lotes de observações seja próxima de zero. Ao realizar a análise de cada resultados utilizando o método de Loteamento, verificamos que a medida com maior intervalo entre observações é a medida de tempo de viagem do servidor 3 para atender chamadas do tipo 2: $\overline{TU}_3^{[2]}$ . Desta forma, o tempo transiente obtido é de 8.000 minutos e o tempo total de simulação é de 5,178 x$10^6$ minutos. De acordo com a análise dos resultados por meio do *Output Analyzer* do software *Arena*, este tempo corresponde ao tempo necessário que gera um suficiente número de observações de cada medida para calcular o intervalo de confiança e evitar correlação. O tempo computacional necessário para rodar o modelo de simulação foi de 0,58 minutos.

A seguir comparamos os resultados de algumas medidas de desempenho obtidos pelo modelo com os obtidos pela simulação utilizando o intervalo de confiança (COSTA NETO, 1977; KELTON et al., 2002) calculado para cada resultado da simulação. A tabela 6.23 apresenta os resultados do modelo hipercubo, os valores médios e o intervalo de confiança dos resultados da simulação para as medidas: tempo médio de viagem de cada servidor para atender chamadas tipo 1 ($\overline{TU}_j^{[1]}$), tempo médio de viagem de cada servidor para atender chamadas tipo 2 ($\overline{TU}_j^{[2]}$), tempo médio de viagem de cada servidor para atender chamadas de qualquer tipo como primeiro a chegar no local ($\overline{TU_j}$) e cargas de trabalho dos servidores.

Tabela 6.23 – Resultados do tempo médio de viagem e carga de trabalho dos servidores

| **Serv.** $j$ | **Modelo** | $\overline{TU}_j^{[1]}$ **tipo 1** | $\overline{TU}_j^{[2]}$ **tipo 2** | $\overline{TU_j}$ **geral** | $\rho_j$ |
|---|---|---|---|---|---|
| SAU 1 | Hipercubo | 5,993 | 10,460 | 6,232 | 0,0578 |
| | Simulação | 6,012 | 10,283 | 6,241 | 0,0594 |
| | (Intev. Conf.) | 5,915 – 6,107 | 10,039 – 10,526 | 6,155 – 6,327 | 0,0576– 0,0612 |
| SAU 2 | Hipercubo | 7,342 | 6,862 | 7,300 | 0,0537 |
| | Simulação | 7,316 | 6,830 | 7,274 | 0,0545 |
| | (Intev. Conf.) | 7,230 – 7,402 | 6,669 – 6,991 | 7,105 – 7,443 | 0,0528 – ,0562 |
| SAU 3 | Hipercubo | 6,686 | 18,450 | 7,003 | 0,0186 |
| | Simulação | 6,683 | 18,450 | 7,041 | 0,0191 |
| | (Intev. Conf.) | 6,810 – 6,556 | - | 6,832 – 7,250 | 0,0178 –0,0204 |
| SAU 4 | Hipercubo | 6,705 | 15,250 | 6,716 | 0,0253 |
| | Simulação | 6,778 | 15,250 | 6,797 | 0,0253 |
| | (Intev. Conf.) | 6,607 – 6,948 | - | 6,628 – 6,966 | 0,0236 –0,0269 |
| SAU 5 | Hipercubo | 3,682 | 4,600 | 3,765 | 0,0185 |
| | Simulação | 3,711 | 4,600 | 3,792 | 0,0182 |
| | (Intev. Conf.) | 3,627 – 3,794 | - | 3,719 – 3,865 | 0,0173 – 0,0191 |

Os resultados da tabela 6.23 mostram que os resultados obtidos pelo modelo estão dentro do intervalo de confiança calculado para os resultados da simulação. Note que estes desvios são pequenos dado que as mesmas matrizes de tempos médios de viagem (tabelas 6.14 e 6.15) utilizadas no modelo, são também dados de entrada da simulação. Em particular, na tabela 6.23, os valores de $\overline{TU}_j^{[2]}$ para os servidores 3, 4 e 5 apresentam intervalos de confiança nulos, pois estes servidores atendem apenas um átomo $i$ com chamadas tipo 2, e portanto o valor de $\overline{TU}_j^{[2]}$ corresponde aos valores de $t_{ji}^{[2]}$ na tabela 6.15 (i.e, são constantes).

Outras medidas agregadas de tempo de viagem também podem ser comparadas, e os resultados obtidos são mostrados na tabela 6.24. Esta tabela apresenta de forma resumida os resultados do modelo comparados com os resultados da amostra, de acordo com a descrição acima. Na quinta e sexta colunas da tabela 6.24, encontram-se o valor médio e o intervalo de confiança para cada medida obtida pela simulação.

Tabela 6.24 – Medidas agregadas de tempo médio de viagem (em minutos)

| Medida | Hipercubo | Amostra | Desvio | Simulação média | Simulação (Intev. Conf.) |
|---|---|---|---|---|---|
| $\overline{T}^{[1]}$ | 6,277 | 5,962 | 5,3% | 6,279 | 6,235 – 6,322 |
| $\overline{T}^{[2]}$ | 8,186 | | | 8,133 | 7,858 – 8,408 |
| $\overline{T}^{,[2]}$ | 7,776 | 7,201 | 7,9% | 7,821 | 7,567 – 8,074 |
| $\overline{T}_t^{,}$ | 25,130 | 25,144 | 0,06% | | |
| $\overline{T}_F$ | 7,776 | 7,771 | 0,06% | 7,821 | 7,567 – 8,074 |
| $\overline{T}_S$ | 17,373 | 17,944 | 3,3% | 17,426 | 17,245– 17,607 |

Com base na comparação dos resultados do modelo com os da amostra verificamos que os desvios são relativamente pequenos. Além disso, o modelo também é validado pelo modelo de simulação, considerando que os resultados obtidos pelo modelo Hipercubo estão dentro do intervalo de confiança dos resultados da simulação apresentados na sexta coluna da tabela 6.24.

## 6.3 Modelo Hipercubo múltiplo despacho modificado para o SAE *Centrovias* com chamadas tipo 1a .

Dados de entrada:

Os dados de entrada para o modelo múltiplo despacho aplicado na seção 6.2, deve ser modificado de forma a diferenciar as chamadas do tipo 1a e o tempo médio de atendimento de cada servidor para chamadas do tipo 1a.

Processo de chegada:

A tabela 6.25 apresenta as taxas de chegada de cada tipo de chamada (1,2 e 1a) em cada átomo do sistema.

Tabela 6.25 – Taxa de chegada de cada tipo de chamada no sistema

| **Atomo** $i$ | **Tipo 1** $\lambda_i^{[1]}$ **cham/h** | **Tipo 2** $\lambda_i^{[2]}$ **cham/h** | **Tipo 1a** $\lambda_i^{[1a]}$ **cham/h** | **Primeiro servidor** | **Segundo servidor** |
|---|---|---|---|---|---|
| **1** | 0,04520 | 0,00340 | 0,01660 | 1 | 2 |
| **2** | 0,04105 | 0,00476 | 0,00714 | 2 | 1 |
| **3** | 0,00693 | 0,00040 | - | 1 | 2 |
| **4** | 0,00911 | 0,00042 | 0,00085 | 3 | 1 |
| **5** | 0.00300 | - | 0,00194 | 3 | 2 |
| **6** | 0.00932 | - | 0,00207 | 4 | 3 |
| **7** | 0.01148 | - | 0,00292 | 4 | 5 |
| **8** | 0.01663 | 0,00196 | 0,00274 | 5 | 4 |

Note que, $\lambda = \sum_{i=1}^{8}(\lambda_i^{[1]} + \lambda_i^{[2]} + \lambda_i^{[1a]}) = 0{,}1878$, como na seção 6.2 (tabela 6.12).

Processo de atendimento:

Como o tempo de atendimento de chamadas do tipo 1a difere do tempo de atendimento de chamadas de emergência (tipo 1 e 2), a tabela 6.26 mostra as duas taxas de atendimento: $\mu_j^{[I]}$ (tipo 1 e 2) e $\mu_j^{[II]}$ (tipo 1a) para cada servidor $j$ do sistema.

Tabela 6.26 – Taxa de atendimento de cada servidor

| **Servidor** $j$ | Atendimento Tipo 1 - $\mu_j^{[I]}$ | Atendimento Tipo 2 - $\mu_j^{[II]}$ |
|---|---|---|
| SAU 1 | 1,1340 | 2,9911 |
| SAU 2 | 1,0387 | 1,2931 |
| SAU 3 | 0,7960 | 1,0026 |
| SAU 4 | 1,0782 | 1,1219 |
| SAU 5 | 1,1213 | 1,5267 |

Os dados de tempos médios de viagem de único despacho de cada servidor a cada átomo é reapresentada na tabela 6.27. Note que, comparando com a tabela 6.14, o tempo médio de viagem de alguns servidores aos seus átomos preferenciais aumenta, pois não estamos considerando os tempos de viagem iguais a zero, dos atendimentos tipo 1a como na análise da seção 6.2.

Tabela 6.27 – Tempo de viagem servidor – átomo $t_{ji}^{[1]}$ (minutos)

| $t_{ji}^{[1]}$ | **Átomo 1** | **Átomo 2** | **Átomo 3** | **Átomo 4** | **Átomo 5** | **Átomo 6** | **Átomo 7** | **Átomo 8** |
|---|---|---|---|---|---|---|---|---|
| SAU 1 | 6,53 | 15,99 | 13,21 | 22,30 | - | - | - | - |
| SAU 2 | 16,18 | 7,65 | 23,18 | - | - | - | - | - |
| SAU 3 | - | - | - | 7,04 | 8,63 | 24,11 | - | - |
| SAU 4 | - | - | - | - | 40,92 | 8,73 | 7,38 | 19,35 |
| SAU 5 | - | - | - | - | - | - | 13,81 | 3,60 |

Como o sistema tem N = 5 servidores, então há $3^5 = 243$ possíveis estados para o sistema. Calculando as probabilidades de estado, obtemos as principais medidas de desempenho do sistema. Apresentamos a seguir algumas das medidas do sistema que sofrem alterações ao consideramos chamadas tipo 1a no sistema. Como nas análises das seções 6.1 e 6.2, algumas medidas obtidas pelo modelo foram comparadas com os resultados da análise da amostra de dados e da simulação do sistema considerando chamadas tipo 1a.

Cargas de trabalho:

A tabela 6.28 apresenta os resultados para cargas de trabalho ($\rho_j^{[I]}$ e $\rho_j^{[II]}$) para cada servidor, discutidas na seção 4.3 do capítulo 4. A última coluna apresenta os resultados das cargas de trabalho calculadas pelo modelo anterior aplicado na seção 6.2.

Tabela 6.28 – Resultados para carga de trabalho dos servidores

| **Servidor *j*** | $\rho_j^{[I]}$ **modelo** | $\rho_j^{[I]}$ **amostra** | $\rho_j^{[II]}$ **modelo** | $\rho_j^{[II]}$ **amostra** | $\rho_j$ **anterior** |
|---|---|---|---|---|---|
| SAU 1 | 0,0525 | 0,0532 | 0,0052 | 0,0055 | 0,0578 |
| SAU 2 | 0,0477 | 0,0478 | 0,0052 | 0,0053 | 0,0537 |
| SAU 3 | 0,0157 | 0,0158 | 0,0027 | 0,0025 | 0,0186 |
| SAU 4 | 0,0209 | 0,0211 | 0,0043 | 0,0041 | 0,0253 |
| SAU 5 | 0,0165 | 0,0165 | 0,0018 | 0,0018 | 0,0185 |

Note que como esperado, os resultados do modelo anterior $\rho_j$ são, aproximadamente, a soma dos resultados das colunas $\rho_j^{[I]}$ e $\rho_j^{[I]}$(modelo) .

Probabilidade de Perda:

A probabilidade de perda para o sistema ($P_p$) é 0,0048, e a probabilidade de perda para chamadas de emergência ($P_p^{[1]} + P_p^{[2]}$) é igual à 0,0139 e para chamadas tipo 1a ($P_p^{[1a]}$) é igual à 0,0067. Estas medidas foram calculadas de forma similar ao exemplo 5 do capítulo 4 (i.e, expressões (4.22) à (4.24)).

Freqüências de despacho:

A tabela 6.29 apresenta as freqüências de despacho de cada servidor a cada átomo, considerando todos os tipos de chamada obtidas pela expressão (4.56) (capitulo 4).

Tabela 6.29 – Freqüência de despacho tipo 1 servidor $j$ – átomo $i$ ( $f_{ji}$ )

| $f_{ji}$ | **Átomo 1** | **Átomo 2** | **Átomo 3** | **Átomo 4** | **Átomo 5** | **Átomo 6** | **Átomo 7** | **Átomo 8** |
|---|---|---|---|---|---|---|---|---|
| SAU 1 | 0,328 | 0,011 | 0,037 | 0,001 | - | - | - | - |
| SAU 2 | 0,013 | 0,268 | 0,002 | - | - | - | - | - |
| SAU 3 | - | - | - | 0,054 | 0,026 | 0,001 | - | - |
| SAU 4 | - | - | - | - | 0,0005 | 0,059 | 0,075 | 0,002 |
| SAU 5 | - | - | - | - | - | - | 0,001 | 0,112 |

Verifique que, se somarmos as freqüências da tabela 6.29, temos que $\sum_{j}^{5}\sum_{i}^{8} f_{ji} = 1$.

Tempo de viagem:

Como os dados de entrada para chamadas do tipo 2 são os mesmos do sistema analisado na seção 6.2, os resultados de tempos de viagem para chamadas tipo 2 sofrem poucas alterações (relacionadas às modificações nas probabilidades de estado) com relação aos resultados daquela análise. Apresentamos a seguir medidas de desempenho relacionadas apenas a único despacho de ambulâncias.

O tempo médio para chamadas do tipo 1 no sistema, calculado pela expressão (4.37) do capítulo 4, substituindo $t_{ji}$ por $t_{ji}^{[1]}$, é : $\overline{T}^{[1]}$ = 7,376 minutos. O resultado para esta medida obtido pela análise da amostra é 7,196 minutos (desvio de -2,5%) e pelo modelo de simulação é 7,465 (desvio de apenas 1,19%). Lembre-se que no sistema anterior, $\overline{T}^{[1]}$= 6,277 (14,89% menor do que o modelo atual).

O tempo médio de viagem para cada servidor considerando chamadas tipo 1 ($\overline{TU}_j^{[1]}$) é calculado pela expressão (4.38) do capítulo 4. Os valores desta medida são apresentados na tabela 6.30, e comparados com os resultados da análise da amostra, a análise anterior da seção 6.2 e resultados da simulação (desvio relativo do modelo com relação a estes).

Tabela 6.30 – Tempo de médio viagem servidor – único despacho $\overline{TU}_j^{[1]}$ (minutos)

| **Servidor** $j$ | $\overline{TU}_j^{[1]}$ **atual (c/tipo 1a)** | $\overline{TU}_j^{[1]}$ **amostra** | **Desvio** | $\overline{TU}_j^{[1]}$ **anterior** | $\overline{TU}_j$ **(min) – simulação** | **Simulação (Intev. Conf.)** |
|---|---|---|---|---|---|---|
| SAU 1 | 7,481 | 7,503 | 0,29% | 5,993 | 7,622 | 7,405-7,839 |
| SAU 2 | 8,157 | 7,871 | -3,6% | 7,342 | 8,215 | 8,123-8,300 |
| SAU 3 | 7,822 | 8,068 | 3,04% | 6,686 | 7,689 | 7,499-7,876 |
| SAU 4 | 8,144 | 8,042 | 1,27% | 6,705 | 8,206 | 8,079-8,332 |
| SAU 5 | 3,724 | 3,701 | -0,62% | 3,682 | 3,706 | 3,627-3,785 |
| $\overline{T}^{[1]}$ | 7,376 | 7,196 | -2,50% | 6,277 | | |

Note na quinta e sexta colunas da tabela 6.30 que o tempo médio de viagem de cada servidor aumenta significantemente com relação aos resultados desta medida obtidos na análise anterior (seção 6.2 - tabelas 6.22 e 6.23). Por exemplo para SAU1, o desvio do modelo anterior para o modelo atual é de 24,83%. Isto se deve ao fato de que, a matriz do tempo de viagem considerada no modelo Hipercubo e na análise da amostra com chamadas tipo 1a (análises desta seção) não considera os tempos de viagem iguais à zero ($t_{ji}^{[1]} = 0$), que correspondem aos atendimentos realizados na base. O modelo Hipercubo e a análise da amostra da seção 6.2 incluem estes dados ($t_{ji}^{[1]} = 0$) no cálculo da matriz dos tempos de viagem (tabela 6.14), pois não diferenciam chamadas tipo 1 e tipo 1a com tempos de atendimento $\mu_j^{[I]}$ e $\mu_j^{[II]}$, respectivamente. Estas diferenças podem também ser verificadas comparando as tabelas 6.14 e 6.27. Desta forma, apesar do modelo atual ser mais caro em termos de esforço computacional, pois o espaço de estados cresce de $O(2^N)$ para $O(3^N)$, o mesmo representa melhor o sistema real analisado. As duas últimas colunas da tabela 6.31 mostram que o modelo Hipercubo é validado pela simulação dado que os resultados para cada medida obtidos pelo modelo estão dentro do intervalo de confiança da respectiva medida obtida pela simulação.

## 6.4 Modelo Hipercubo para análise do SAE *Centrovias* 2 – carro médico:

Coleta de Dados

Os dados da operação e configuração deste novo sistema foram coletados durante o período de janeiro à setembro de 2004, e de forma similar a coleta de dados do sistema anterior. Como na análise de dados da seção 6.2, durante este período não foram constatados períodos de pico ou significantes variações ao longo dos meses analisados. O número total de eventos ocorridos durante o período é 1.498 eventos.

Divisão dos trechos de rodovia em átomos:

A divisão do trecho em átomos corresponde a mesma divisão de 8 átomos descrita na seção 6.2 e na figura 6.3. Como discutido anteriormente na seção 4.4 do capítulo 4, a aplicação do modelo hipercubo múltiplo despacho para um sistema com servidores diferenciados requer que os átomos do sistema sejam subdivididos em camadas de acordo com o tipo de chamada que possui uma lista de despacho particular. A tabela 6.31 apresenta como os 8 átomos são subdivididos de acordo com o tipo de chamada e a lista de preferência de despacho para cada sub-átomo.

Tabela 6.31 Tipo de chamada em cada sub-átomo e lista de preferência de cada átomo

| **Átomo e sub-átomos** | **chamadas** | **primeiro** | **segundo** | **terceiro** |
|---|---|---|---|---|
| 1a | 1, 2a | 2 | 3 | - |
| 1b | 1, 2b, 3 | 1 | 2 | 3 |
| 2a | 1, 2a | 3 | 2 | - |
| 2b | 2b,3 | 1 | 3 | 2 |
| 3a | 1,2a | 2 | 4 | - |
| 3b | 2b, 3 | 1 | 2 | 4 |
| 4a | 1,2a | 4 | 2 | - |
| 4b | 2b | 1 | 4 | 2 |
| 5 | 1,2a | 4 | 5 | 6 |
| 6a | 1,2a | 5 | 4 | - |
| 6b | 2b | 4 | 5 | 6 |
| 7 | 1,2a, 2b | 5 | 6 | 4 |
| 8 | 1,2a | 6 | 5 | |

Note que, chamadas do tipo 1 e 2a possuem 2 servidores em sua lista de despacho, e chamadas dos tipos 2b e 3 possuem até 3 servidores possíveis. Não foi necessário subdividir

os átomos 5, 7 e 8 em camadas, dado que todos os tipos de chamada nestes átomos possuem uma mesma lista de despacho.

**Análise Estatística dos dados:**

Processo de Chegada

A tabela 6.32, a seguir, apresenta os resultados obtidos da análise do processo de chegada de chamadas em cada átomo do sistema, o número de observações em cada átomo e a taxa de chegada total $\lambda_i$ em cada átomo *i* (chamadas/hora), igual à $\sum_{m=1} \lambda_i^m$, onde $\lambda_i^m$ corresponde a taxa de chegada de cada tipo de chamada *m* = 1, 2a, 2b e 3 no átomo *i*.

Tabela 6.32 – Dados do processo de chegada

| **Átomo** *i* | **Número de eventos** | **Taxa de chegada** $\lambda_i$ **(cham/h)** | **Sub atomo** | **Chamadas Tipo 1** $\lambda_i^{[1]}$ | **Chamadas Tipo 2a** $\lambda_i^{[2a]}$ | **Chamadas Tipo 2b** $\lambda_i^{[2b]}$ | **Chamadas Tipo 3** $\lambda_i^{[3]}$ |
|---|---|---|---|---|---|---|---|
| **1** | 550 | 0,0848 | 1a<br>1b | 0,0328<br>0,0308 | 0,00090 | 0,01967 | 0,00060 |
| **2** | 399 | 0,0579 | 2a<br>2b | 0,0528 | 0,00145 | 0,00252 | 0,00118 |
| **3** | 34 | 0,0062 | 3a<br>3b | 0,0036 | 0,00019 | 0,00228 | 0,00017 |
| **4** | 67 | 0,0100 | 4a<br>4b | 0,0091 | 0.00029 | 0,00062 | |
| **5** | 57 | 0,0085 | | 0,0084 | 0,00015 | | |
| **6** | 133 | 0,0194 | 6a<br>6b | 0,0185 | 0,00073 | 0,00016 | |
| **7** | 104 | 0,0153 | | 0,0146 | 0,00059 | 0,00015 | |
| **8** | 154 | 0,0230 | | 0,0213 | 0,00164 | - | |

A taxa total de chegadas no sistemas é $\lambda = 0{,}2251$ chamadas/hora. O coeficiente de variação (razão entre o desvio padrão e a média) dos intervalos entre chegadas também foi calculado para cada átomo do sistema. Verificou-se que este fica muito próximo de 1, o que sugere que o processo de chegada possa ser aproximado por uma distribuição Poisson.

Teste de aderência: Como na análise de dados do sistema *Centrovias* anterior (seção 6.2), realizamos análises estatísticas para verificar se o processo de chegada é Poisson. Realizando os testes *Kolmogorov-Smirnov, Anderson Darling* e *Qui-quadrado,* os resultados obtidos mostraram que, com nível de significância $\alpha = 0{,}05$ e para todos os átomos, não se pode

rejeitar a hipótese de que os intervalos entre chegadas sucessivas tem distribuição exponencial.

Processo de Atendimento

A tabela 6.33, a seguir, apresenta os resultados obtidos da análise do processo de atendimento realizado por cada servidor do sistema, considerando que o servidor 1 corresponde ao carro médico e os demais as 5 viaturas resgates. Estes dados são: tempo médio de atendimento em minutos, desvio padrão, o coeficiente de variação (razão entre o desvio padrão e a média), taxa média de atendimento e fração de ocupação (obtida através da razão entre a soma total do tempo de atendimento do servidor e o tempo total de operação). O valor de $\mu$ é 6,3714. Portanto, $\rho = \frac{\lambda}{\mu} = 0{,}0328$.

Tabela 6.33 – Dados do processo de atendimento

| **Átomo** $i$ | **Número de observações** | **Tempo atendimento (horas)** | **Desvio padrão (minutos)** | **Coef. de variação** | **Taxa** $\mu_j$ **(cham/h)** | **Fração de ocupação** |
|---|---|---|---|---|---|---|
| **1** | 366 | 0,8240 | 1,0765 | 1,3064 | 1,2136 | 0,0442 |
| **2** | 426 | 0,9808 | 0,7717 | 0,7868 | 1,0196 | 0,0613 |
| **3** | 395 | 0,9687 | 0,5913 | 0,6104 | 1,0323 | 0,0561 |
| **4** | 142 | 1,1011 | 0,6856 | 0,6226 | 0,9082 | 0,0229 |
| **5** | 245 | 0,9317 | 0,5913 | 0,6346 | 1,0733 | 0,0335 |
| **6** | 165 | 0,8894 | 0,5384 | 0,6053 | 1,1244 | 0,0215 |

Note que os coeficientes de variação dos servidores 2 -6 (coluna 5 da tabela 6.33) não são tão próximos de 1.0, o que sugere que o processo de atendimento não deve ser exponencial. Note também que os servidores 1, 2 e 3 (localizados na rodovia Washington Luiz) possuem maior carga de trabalho, dado que suas áreas de atendimento correspondem ás áreas com maior taxa de chegadas (veja também a tabela 6.32).

Como na análise da seção 6.2, verificamos a variabilidade nos tempos de atendimento entre os servidores, por meio da análise de variância ANOVA com nível de significância $\alpha = 0{,}05$. Os resultados mostraram que, como no sistema *Centrovias* anterior (seção 6.2), as diferenças entre as médias dos tempos de atendimento entre os servidores são significantes e estes devem ser considerados não homogêneos (i.e, as taxas de atendimento $\mu_j$ são distintas) na análise.

Teste de aderência: Como no processo de chegada, também foram realizadas análises estatísticas para verificar se a distribuição dos tempos de atendimento pode ser considerada exponencial. Os resultados obtidos nesta análise mostraram que esta hipótese deve ser rejeitada, com nível de significância $\alpha = 0{,}05$. Este resultado também foi obtido nos casos anteriores (*Anjos do Asfalto* e *Centrovias 1*), mas o sistema pode ser analisado aproximadamente pelo modelo Hipercubo sem que a análise seja comprometida (conforme discussão na seção 3.3 do capítulo 3).

Matriz dos Tempos de viagem:

Para o cálculo dos tempos de viagem, utilizamos os dados dos instantes de acionamento dos servidores e instantes de chegada no local da chamada. Utilizando o intervalo médio entre estas duas medidas, obtemos a matriz dos tempos de viagem de cada servidor a cada átomo (no caso átomos em que o servidor pode atender), considerando que há dois tipos de despacho: chamadas único despacho (tipo 1) e chamadas múltiplo despacho (tipos 2a, 2b e 3). Desta forma, a tabela 6.34 mostra a matriz do tempo médio de viagem servidor – átomo para chamadas único despacho ($t_{ji}^{[m=1]}$) e a tabela 6.35 apresenta esta matriz para chamadas múltiplo despacho ($t_{ji}^{[m>1]}$).

Note que, como discutido na seção 6.2, $t_{ji}^{[m=1]} \neq t_{ji}^{[m>1]}$, pois os tempos médios de cada servidor a cada átomo para chamadas único e múltiplo apresentam significante variabilidade para alguns servidores. Por exemplo, veja na tabela 6.36 que, no caso do carro médico, grande parte das chamadas tipo 1 são atendidas na própria base (i.e, $t_{ji} = 0$), quando um usuário da rodovia estaciona no SAU, pedindo assistência médica.

Tabela 6.34 – Tempo de viagem servidor – átomo $t_{ji}^{[1]}$ (minutos) – único despacho

| $t_{ji}^{[t=1]}$ servidor | **Átomo 1** | **Átomo 2** | **Átomo 3** | **Átomo 4** | **Átomo 5** | **Átomo 6** | **Átomo 7** | **Átomo 8** |
|---|---|---|---|---|---|---|---|---|
| **1** | 2,67 | - | - | - | - | - | - | - |
| **2** | 7,05 | 15,29 | 11,62 | *14,03 | - | - | - | - |
| **3** | 13,83 | 7,51 | - | - | - | - | - | - |
| **4** | - | - | 12,48 | 8,10 | 7,01 | 11,66 | - | - |
| **5** | - | - | - | **7,15 | **6,03 | 11,52 | 5,05 | 13,17 |
| **6** | | | | | | 16,73 | 11,43 | 6,24 |

* calculados com base na distância do servidor ao centróide do átomo; ** apenas uma observação.

Tabela 6.35 – Tempo de viagem servidor – átomo $t_{ji}^{[t>1]}$ (minutos) – múltiplo despacho

| $t_{ji}^{[t>1]}$ | **Átomo 1** | **Átomo 2** | **Átomo 3** | **Átomo 4** | **Átomo 5** | **Átomo 6** | **Átomo 7** | **Átomo 8** |
|---|---|---|---|---|---|---|---|---|
| **1** | 6,76 | 15,27 | 11,30 | 23,49 | - | - | - | - |
| **2** | 6,72 | 13,96 | 10,24 | **14,03 | - | - | - | - |
| **3** | 14,20 | 6,81 | - | - | - | - | - | - |
| **4** | - | - | 18,10 | 13,18 | 4,54 | 23,53 | 37,8 | 37,8 |
| **5** | - | - | - | 7,15 | 16,24 | 17,14 | 8,37 | 10,43 |
| **6** | - | - | - | - | **9,47 | 16,73 | 9,22 | 6,69 |

** apenas uma observação.

A tabela 6.36 a seguir mostra a localização (km na rodovia) e proporção de chamadas atendidas na base (tempo de viagem nulo) de cada servidor.

Tabela 6.36 –Localização e fração de atendimento na base do servidor

| **Servidor** *j* | **Localização** | **Fração de atendimento na base** |
|---|---|---|
| **1** | Km 217-SP310 | 34,61% |
| **2** | Km 217-SP310 | 12,44% |
| **3** | Km 187-SP310 | 8,86% |
| **4** | Km 144-SP225 | 14,08% |
| **5** | Km 199-SP225 | 11,84% |
| **6** | Km 224-SP225 | 11,51% |

**Aplicação do Modelo Hipercubo:**

O modelo Hipercubo de múltiplo despacho para SAEs em rodovias considerando servidores diferenciados (p.e, carro médico e as viaturas resgates), discutido no capítulo 4, foi implementado para analisar o estudo de caso do novo sistema de atendimento médico da concessionária *Centrovias*. Os resultados obtidos para as algumas das principais medidas de desempenho são apresentados a seguir.

Equações de Equilíbrio:

As equações de equilíbrio são determinadas de acordo com a discussão do exemplo 6 do capítulo 4, de um SAE em rodovias com política de múltiplo despacho e servidores diferenciados. Neste novo sistema *Centrovias* (*Centrovias 2*) há 6 servidores (1 carro médico e 5 carros resgates), então há $2^6 = 64$ estados possíveis do sistema (como no sistema *Anjos do Asfalto*).

Analisando, por exemplo, o estado B = {11001}, com base na lista de preferência de despacho da tabela 6.31, temos a seguinte equação de equilíbrio:

$$p_{111000}(\lambda - [\lambda_{1a}^{[1]} + \lambda_{1a}^{[2a]} + \lambda_{1b}^{[1]} + \lambda_{1b}^{[2b]} + \lambda_{1b}^{[3]} + \lambda_{2a}^{[1]} + \lambda_{2a}^{[2a]} + \lambda_{2b}^{[2b]} + \lambda_{2b}^{[3]}] + \mu_1 + \mu_2 + \mu_3) =$$
$$p_{000000}(\lambda_{1b}^{[3]} + \lambda_{2b}^{[3]}) + p_{001000}(\lambda_{1b}^{[2b]} + \lambda_{1b}^{[3]} + \lambda_{2b}^{[2b]} + \lambda_{2b}^{[3]} + \lambda_{3b}^{[2b]} + \lambda_{3b}^{[3]}) + p_{010000}(\lambda_{1b}^{[2b]} + \lambda_{1b}^{[3]} + \lambda_{2b}^{[2b]}) +$$
$$p_{100000}(\lambda_{1a}^{[2a]} + \lambda_{1b}^{[2b]} + \lambda_{1b}^{[3]} + \lambda_{2a}^{[2a]} + \lambda_{2b}^{[2b]} + \lambda_{2b}^{[3]} + \lambda_{3b}^{[3]}) + p_{011000}(\lambda_{1b}^{[1]} + \lambda_{1b}^{[2b]} + \lambda_{1b}^{[3]} + \lambda_{2b}^{[2b]} + \lambda_{2b}^{[3]})$$
$$+ p_{101000}(\lambda_{1a}^{[1]} + \lambda_{1a}^{[2a]} + \lambda_{1b}^{[1]} + \lambda_{1b}^{[2b]} + \lambda_{1b}^{[3]} + \lambda_{2a}^{[1]} + \lambda_{2a}^{[2a]} + \lambda_{2b}^{[2b]} + \lambda_{2b}^{[3]} + \lambda_{3a}^{[1]})$$
$$p_{110000}(\lambda_{1a}^{[1]} + \lambda_{1a}^{[2a]} + \lambda_{1b}^{[1]} + \lambda_{1b}^{[2b]} + \lambda_{1b}^{[3]} + \lambda_{2a}^{[1]} + \lambda_{2a}^{[2a]} + \lambda_{2b}^{[2b]} + \lambda_{2b}^{[3]}) + p_{111010}(\mu_5) + p_{111001}(\mu_6)$$

(6.9)

Observe que, de acordo com o lado esquerdo da equação (6.9), o sistema não permite que chamadas de qualquer tipo nos átomos 1 e 2 sejam atendidas, pois neste estado os três servidores preferenciais destes átomos (1, 2 e 3) estão ocupados.

Com relação ao lado direito da equação (6.9), temos por exemplo, que as transições para dentro do estado B = {111000} envolvendo múltiplo despacho de ambulâncias são:

{000000} $\rightarrow$ {111000} – que ocorre com a chegada de uma chamada tipo 3 nos átomos 1 ou 2 (taxa total $\lambda_{1b}^{[3]} + \lambda_{2b}^{[3]}$), atendidas pelo carro médico e 2 resgates;

{001000} $\rightarrow$ {111000}- que ocorre com a chegada de uma chamada tipo 2b ou tipo 3 nos átomos 1, 2 ou 3 (note que, podem ser atendidas também pelo terceiro servidor, se está estiver livre e os dois primeiros estão ocupados)

{010000} $\rightarrow$ {111000}- que ocorre com a chegada de uma chamada tipo 2b nos átomos 1 ou 2, ou uma chamada do tipo 3 nos átomos 1 ou 2 (taxa total $\lambda_{1b}^{[2b]} + \lambda_{1b}^{[3]} + \lambda_{2b}^{[2b]}$);

{100000} $\rightarrow$ {111000}- que ocorre com a chegada de uma chamada tipo 2a ou 2b nos átomos 1 ou 2, ou uma chamada do tipo 3 nos átomos 1, 2 ou 3;

**Resultados para algumas medidas de desempenho obtidas pelo modelo:**

Determinando as probabilidades de estados do sistema, é possível calcular algumas medidas preliminares de desempenho do sistema (discutidas nas seções 4.2 e 4.4 para os exemplos 4 e 6 do capítulo 4) e compara-las com os resultados da amostra de dados. Similarmente, a análise da seção 6.2, os resultados obtidos para as probabilidades de estados do sistema mostraram que o mesmo tem probabilidade muito pequena de estar saturado (todos servidores ocupados),

dado que $p_{\{111111\}} = 0{,}0000002$. Na maior parte do tempo o sistema permanece ocioso (todos servidores livres), pois $p_{\{000000\}} = 0{,}7964$.

Carga de trabalho dos servidores

Os resultados para as cargas dos servidores do sistema são apresentados na tabela 6.37. Na coluna 3 desta tabela estão a fração de ocupação obtida da amostra de dados, e a coluna 4 apresenta o desvio relativo para resultado do modelo com relação a amostra.

Tabela 6.37 – Resultados para carga de trabalho dos servidores

| **Servidor** | $\rho_j$ **Modelo** | $\rho_j$ **Amostra** | **Desvio relativo** |
|---|---|---|---|
| **1** | 0,0454 | 0,0442 | 2,71% |
| **2** | 0,0621 | 0,0613 | 1,30% |
| **3** | 0,0576 | 0,0561 | 2,67% |
| **4** | 0,0226 | 0,0229 | -1,31% |
| **5** | 0,0336 | 0,0335 | 0,29% |
| **6** | 0,0211 | 0,0215 | -1,86% |

Note que os desvios dos resultados do modelo para a amostra são pequenos.

Probabilidade de Perda:

A probabilidade de perda para chamadas tipo 1 ( único despacho - $P_p^{[1]}$ ) é igual à 0,0063, para chamadas tipo 2 (2a e 2b – duplo despacho $P_p^{[2]}$) é igual à 0,0023 e para chamadas tipo 3 (triplo despacho - $P_p^{[3]}$ ) é igual à 0,0019. A probabilidade de perda para o sistema ($P_p$) é 0,00572. Estas medidas foram calculadas de forma similar aos exemplo 4 e 6 do capítulo 4 (expressão (4.62)).

Freqüências de despacho:

Como discutido no capítulo 4, podemos definir diversas medidas de freqüência de despacho de acordo com cada tipo de chamada no sistema. Apresentamos a seguir os resultados obtidos para apenas algumas destas medidas. Algumas já foram definidas na análise do sistema anterior com múltiplo despacho, tais como: freqüências de despacho para chamadas tipo 1 (simples despacho), freqüências de despacho chamadas tipo 2 que são atendidas por 2

servidores (tipo 2a e 2b), freqüências de despacho chamada tipo 2 que são atendidas por um servidor. Outras medidas adicionais podem também ser determinadas, por exemplo: freqüência de despacho para chamadas tipo 3, freqüência de chamadas para chamadas tipo 3, atendidas por 2 ou 1 servidores, entre outras.

(i) Freqüências de despacho tipo 1:

A tabela 6.38, apresenta as freqüências de despacho do tipo 1 de cada servidor a cada átomo, considerando o total de despachos tipo 1 no sistema, ou seja $f_{ji}^{[1]}$. Estes resultados são obtidos pela expressão (4.25) do capítulo 4.

Tabela 6.38 – Freqüência de despacho tipo 1 servidor $j$ – átomo $i$ ($f_{ji}^{[1]}$)

| $f_{ji}^{[1]}$ | **Átomo 1** | **Átomo 2** | **Átomo 3** | **Átomo 4** | **Átomo 5** | **Átomo 6** | **Átomo 7** | **Átomo 8** |
|---|---|---|---|---|---|---|---|---|
| **1** | 0,1543 | - | - | - | - | - | - | - |
| **2** | 0,1666 | 0,0137 | 0,0178 | 0,0010 | - | - | - | - |
| **3** | 0,0093 | 0,2605 | - | - | - | - | - | - |
| **4** | - | - | 0,0011 | 0,0468 | 0,0430 | 0,0031 | - | - |
| **5** | - | - | - | - | 0,0009 | 0,0939 | 0,0741 | 0,0021 |
| **6** | - | - | - | - | - | - | 0,0024 | 0,1093 |

Verifique que, se somarmos as freqüências da tabela 6.15, temos que $\sum_{j}^{6}\sum_{i}^{8} f_{ji}^{[1]} = 1$.

(ii) Freqüências de despacho tipo 2:

As freqüências de chamadas tipo 2 (2a e 2b) com despacho de dois servidores $f_{(j,k)i}^{[2]}$ para cada átomo $i$ também foram calculadas, onde o servidor $j$ é o servidor preferencial e o servidor $k$ corresponde ao segundo ou terceiro (no caso de tipo 2b) servidor da lista de despacho deste átomo. Estas medidas são obtidas pela expressão (4.28) do capítulo 4, e $f_{(j,k)i}^{[2]}$ é calculada considerando o total de despachos tipo 2 no sistema. Por exemplo para as freqüências de despacho em que $j$=1 (carro médico) temos: $f_{(1,2)1b}^{[2]} = 0{,}5667$, $f_{(1,3)1b}^{[2]} = 0{,}0273$, $f_{(1,2)2b}^{[2]} = 0{,}0047$, $f_{(1,3)2b}^{[2]} = 0{,}0724$, $f_{(1,2)3b}^{[2]} = 0{,}0656$, $f_{(1,2)4b}^{[2]} = 0{,}0004$ e $f_{(1,4)4b}^{[2]} = 0{,}0184$. Os demais valores de $f_{(1,k)i}^{[2]}$ são nulos, de acordo com a lista de preferência de despacho dos átomos do sistema (tabela 6.31).

As medidas de freqüência de despacho para chamadas tipo 2 em que o atendimento é realizado como único despacho (se um dos dois servidores estiver ocupado no caso de chamadas tipo 2a, ou se dois dos três servidores estiverem ocupados no caso de chamadas tipo 2b) também foram obtidas. Como discutido no capítulo 4, a soma das freqüências de despacho tipo 2 é dada por: $\sum_{i=1}^{8}\left[\sum_{j=1}^{6} f_{ji}^{[2]} + \sum_{j=1}^{5}\sum_{k=j+1}^{6} f_{(j,k)i}^{[2]}\right] = 1$, de acordo com a expressão (4.32) do capítulo 4.

Várias medidas adicionais podem também ser calculadas, como por exemplo, as freqüências de triplo despacho no sistema $f_{(j,k,l)i}^{[3]}$, onde *j, k* e *l* correspondem respectivamente, aos servidores 1, 2 e 3 da lista de preferência de despacho do átomo *i*, como descrito no capítulo 4. Considerando os despachos triplos do sistema em que $j$ =1, temos: $f_{(1,2,3)1b}^{[3]}$=0,2645, $f_{(1,3,2)2b}^{[3]} = 0{,}5183$ e $f_{(1,2,4)3b}^{[3]} = 0{,}0796$.

Medidas adicionais de tempo de viagem comentadas no capitulo 4 também podem ser calculadas para descrever este sistema. A seguir, mostramos apenas os resultados para o tempo de viagem de cada servidor para atender chamadas do tipo 1, e comparamos os resultados com a análise da amostra (tabela 6.39). Note que os desvios em relação aos resultados obtidos da amostra são suficientemente pequenos.

Tabela 6.39 – Tempo de médio viagem para cada servidor – único despacho $\overline{TU}_j^{[1]}$ (minutos)

| **Servidor** $j$ | $\overline{TU}_j^{[1]}$ **modelo** | $\overline{TU}_j^{[1]}$ **amostra** | **Desvio** |
|---|---|---|---|
| **1** | 2,674 | 2,674 | 0,0% |
| **2** | 8,060 | 7,868 | 2,44% |
| **3** | 7,729 | 7,562 | 2,21% |
| **4** | 7,771 | 7,895 | -1,57% |
| **5** | 8,707 | 8,898 | -2,15% |
| **6** | 6,356 | 6,602 | -3,73% |

## 6.5 Algoritmo GA/Hipercubo para análise do sistema Anjos do Asfalto:

Esta seção apresenta alguns resultados obtidos pelo Algoritmo Genético combinado com o modelo Hipercubo exato. O algoritmo foi aplicado ao sistema *Anjos do Asfalto*, descrito brevemente na seção 2.2 do capítulo 2. As medidas de desempenho da configuração inicial deste sistema, descritas pelo modelo Hipercubo, foram apresentadas na seção 6.1, deste capítulo. Os principais componentes neste algoritmo genético foram descritos na seção 5.6 do capítulo 5.

### 6.5.1 Configuração Inicial:

A configuração inicial do sistema pode ser representada em um cromossomo de acordo com a representação discutida na seção 5.5.1, que é baseada no tamanho dos átomos do sistema. Utilizando os dados dos tamanhos dos átomos da configuração inicial, apresentados na tabela 6.1, temos o seguinte cromossomo:

| | | | | |
|---|---|---|---|---|
| 0.50 | 0.50 | 0.50 | 0.50 | 0.22 |

A figura 6.1 da seção 6.1 ilustra esta configuração. Note que, com exceção dos dois últimos átomos (9 e 10), cada átomo corresponde a metade da distância entre os servidores (i.e, $y_j = 0{,}5$ ). A tabela 6.40 mostra as três medidas de desempenho em destaque na análise a seguir, conforme descrito na seção 5.5.5: $\overline{T}$ - tempo médio de viagem no sistema (expressões (4.12) e (5.3)); $P_{\bar{tv}>10}$ - fração de chamadas atendidas em tempo superior à 10 minutos (expressão (5.4)) e $\sigma_\rho$ - desvio padrão das cargas de trabalho (expressão (5.6)).

Tabela 6.40– Três medidas de desempenho da configuração original

| **Medidas** | **Conf. original** |
|---|---|
| $\overline{T}$ (min) | 7,9121 |
| $P_{\bar{tv}>10}$ | 0,2995 |
| $\sigma_\rho$ | 0,05507 |

### 6.5.2 Algoritmo Enumerativo/Hipercubo:

Como discutido no capítulo 5, nos procedimentos de geração da população inicial e mutação, consideramos a alternativa de utilizar valores entre 0.2 e 0.8, discretizados por um intervalo Δ. Desta forma, em função do tamanho do presente problema, é possível utilizar um algoritmo enumerativo exaustivo para realizar todas as combinações possíveis destes valores em cada gene, e simplesmente escolher a melhor entre elas. Tal algoritmo pode ser utilizado para avaliar a qualidade das soluções obtidas pelo algoritmo GA/Hipercubo. Como descrito anteriormente, o algoritmo GA/Hipercubo foi testado com três intervalos: Δ = 0.05, Δ = 0.03 e Δ = 0.01. Porém, devido a restrições computacionais, o algoritmo enumerativo foi aplicado apenas para Δ = 0.05 e Δ = 0.03, considerando as três funções de *fitness*: $\overline{T}$ - tempo médio de viagem no sistema (min) (expressão 5.3), $P_{\bar{tv}>10}$ - fração de chamadas atendidas em tempo superior à 10 min (expressão 5.4) e $\sigma_\rho$ - desvio padrão das cargas de trabalho (expressão 5.6)

O número de combinações necessárias para o intervalo Δ = 0.05, é $13^5$, dado que há 5 genes ($N$-1, onde $N$ = 6 servidores) e 13 valores dentro do intervalo (0,1,...,$M = \frac{0.6}{\Delta}$). Similarmente, o número de combinações necessárias para o intervalo Δ = 0.03 , é $21^5$.

### 6.5.3 Resultados obtidos com as três funções *fitness*:

Inicialmente, conduzimos um conjunto de experimentos utilizando individualmente as três funções objetivo (*fitness*) acima: $\overline{T}$, $P_{\bar{tv}>10}$ e $\sigma_\rho$. A seguir são apresentados os seguintes análises com cada função *fitness*:

- Para todos os intervalos Δ = 0.01, 0.03 e 0.05 e para o caso contínuo, foram realizados 2 conjuntos de 20 rodadas do algoritmo,utilizando diferentes sementes para gerar os números aleatórios no algoritmo, com G (número de gerações) igual à 1000 e 2000. Para cada um destes conjuntos, apresentamos a melhor solução, a média das soluções (e o desvio padrão) encontradas nas 20 rodadas, outras medidas de desempenho da melhor solução e número médio de gerações necessárias para encontrar a melhor solução. As tabelas 6.41 a 6.43 apresentam a melhor configuração obtida pelo algoritmo genético em termos das três medidas de funções objetivo testadas individualmente. Para os três casos considerados, o algoritmo

genético encontrou a melhor solução encontrada pelo algoritmo enumerativo. Estas soluções estão apresentadas abaixo.

Função objetivo : Tempo médio de viagem no sistema ( $\overline{T}$ ):

Tabela 6.41 – Três medidas de desempenho da melhor solução encontrada:

| Medidas | Melhor solução obtida | Conf. original | Melhora |
|---|---|---|---|
| $\overline{T}$ (min) | **7,7781** | **7,9121** | **1,69%** |
| $P_{\overline{tv}>10}$ | 0,2753 | 0,2995 | 8,1% |
| $\sigma_\rho$ | 0,05335 | 0,05507 | 3,12% |

A tabela 6.41 mostra que todas as três medidas de desempenho são reduzidas, porém a redução da medida tempo médio de viagem no sistema (função objetivo) é de apenas 1,69%. A figura 6.5 ilustra a melhor configuração encontrada com $\Delta$ = 0.03, cuja combinação de genes é dada pelo cromossomo abaixo:

| 0.41 | 0.44 | 0.44 | 0.41 | 0.29 |
|---|---|---|---|---|

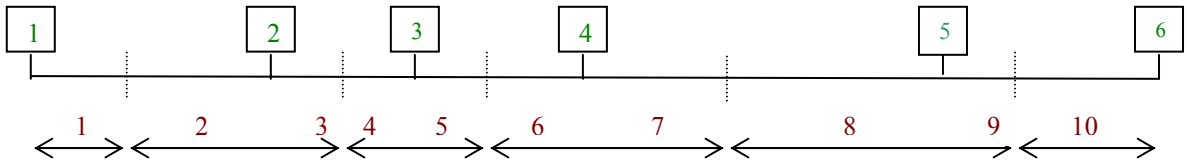


Figura 6.5 – Esquema da distribuição de átomos e servidores ao longo da rodovia no sistema *Anjos do Asfalto* de acordo com a configuração do cromossomo acima

Função objetivo : Fração de chamadas atendidas em tempo superior à 10 minutos ( $P_{\overline{tv}>10}$ ):

Tabela 6.42 – Três medidas de desempenho da melhor solução encontrada:

| Medidas | Melhor solução obtida | Conf. original | Melhora |
|---|---|---|---|
| $\overline{T}$ (min) | 8,041 | 7,9121 | 1,6% |
| $P_{\overline{tv}>10}$ | **0,2545** | **0,2995** | **-17,68%** |
| $\sigma_\rho$ | 0,0511 | 0,05507 | -7,77% |

Note na tabela 6.42 que, a redução obtida para a fração de chamadas atendidas em tempo superior à 10 minutos $P_{\overline{tv}>10}$ (função objetivo) é de 17,68%. Observe também que, nesta

configuração há um melhor balanceamento das cargas de trabalho, pois o seu desvio padrão reduz-se em 7,77%. No entanto, o tempo médio de viagem no sistema aumenta em 1,6%.

Função objetivo : Desvio padrão das cargas de trabalho dos servidores ($\sigma_\rho$):

Tabela 6.43 – Três medidas de desempenho da melhor solução encontrada:

| Medidas | Melhor solução obtida | Conf. original | Melhora |
|---|---|---|---|
| $\overline{T}$ (min) | 8,9446 | 7,9121 | -13,05% |
| $P_{\bar{tv}>10}$ | 0,3781 | 0,2995 | -26,24% |
| $\sigma_\rho$ | **0,02451** | **0,05507** | **55,49%** ← |

Ao observar os resultados da tabela 6.43, notamos que a medida do desvio padrão das cargas de trabalho (função objetivo que pretendemos minimizar) é reduzida significantemente em 55,5%. Por outro lado, note que o tempo médio de viagem no sistema aumenta em 13,05% com relação ao resultado da configuração inicial e a medida da fração de chamadas atendidas em tempo superior a 10 minutos aumenta em 26,24%. A melhor combinação de genes é encontrada para $\Delta = 0.03$, representada pelo cromossomo abaixo e ilustrada na figura 6.6:

| 0.80 | 0.47 | 0.80 | 0.20 | 0.20 |
|---|---|---|---|---|

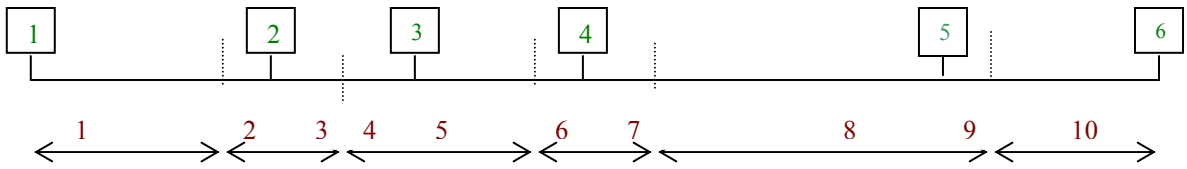


Figura 6.6 – Esquema da distribuição de átomos e servidores ao longo da rodovia no sistema *Anjos do Asfalto* de acordo com a configuração do cromossomo acima

Desta forma, os resultados obtidos com o GA/Hipercubo (ou algoritmo enumerativo/Hipercubo) utilizando as três diferentes funções objetivo mostram que estas medidas analisadas são conflitantes. Por exemplo, a melhor solução obtida no balanceamento das cargas de trabalho resultou em piores valores para o tempo médio de viagem no sistema ($\overline{T}$) e fração de chamadas atendidas em tempo superior a 10 min ($P_{\bar{tv}>10}$), o que não é o objetivo em termos do nível de serviço oferecido ao usuário.

**6.5.4**. **Aplicação do método $\varepsilon$ –restrito de otimização bi-objetivo :**

Como mostrado nos resultados apresentados na tabela 6.43, as diferentes medidas de desempenho do sistema que pretendemos minimizar podem ser conflitantes. Na seção 5.7 propomos avaliar um *trade-off* entre estas duas das medidas conflitantes ($\sigma_\rho, \overline{T}$) durante a busca da configuração ótima do sistema, considerando o problema bi-objetivo. Por exemplo, sendo o primeiro objetivo minimizar o desvio padrão das cargas de trabalho dos servidores $\sigma_\rho$, e o segundo minimizar o tempo médio de viagem no sistema $\overline{T}$. Assim a função objetivo do problema bi-objetivo é: minimizar Z = (*f(x),g(x)*) = ($\sigma_\rho, \overline{T}$).

O método $\varepsilon$ –restrito discutido na seção 5.7 foi utilizado, escolhendo-se como função *fitness* minimizar o desvio padrão das cargas de trabalho dos servidores $\sigma_\rho$, sujeito a restrição da medida tempo médio de viagem ($\overline{T}_{linit}$). Como descrito na seção 5.7, o problema deve ser redefinido da forma:

Minimizar $Z = f(x) = \sigma_\rho$
s.a $g(x) = \overline{T} < \overline{T}_{linit}$
$x \in X^*$, onde $\overline{T}_{linit}$ corresponde ao valor limitante superior de $\overline{T}$.

Uma vez definidos a função objetivo e a restrição do problema, utilizamos os métodos descritos nas seções 6.2 e 6.3 (algoritmo enumerativo e GA/Hipercubo) para prescrever a solução ótima do problema para um dado valor de $\overline{T}_{linit}$. Desta forma conduzimos uma única rodada do algoritmo, inserindo um procedimento para variar o valor de $\overline{T}_{linit}$ de forma a atualizar a curva de soluções eficientes encontrada em função de $\overline{T}_{linit}$ para os dois métodos utilizados (enumerativo e genético). Em todos os experimentos, o algoritmo GA/Hipercubo encontrou a melhor solução obtida pelo algoritmo enumerativo/Hipercubo. As tabelas 6.44 e 6.45, apresentam os resultados das melhores configurações encontradas para cada valor de $\overline{T}_{linit}$ em termos do desvio padrão das cargas de trabalho $\sigma_\rho$ (medida *fitness*) e do tempo médio de viagem no sistema $\overline{T}$, para $\Delta$ = 0.05 e 0.03, respectivamente

Tabela 6.44 – Resultados para $\sigma_\rho$ e $\overline{T}$ - $\Delta = 0.05$

| Restrição $\overline{T}_{linit}$ (min) | *Fitness* $\sigma_\rho$ | $\overline{T}$ (min) |
|---|---|---|
| 7,8 | 0,04678 | 7,796 |
| 8,0 | 0,03387 | 7,988 |
| 8,2 | 0,02967 | 8,199 |
| 8,4 | 0,02737 | 8,381 |
| 8,5 | 0,02632 | 8,484 |
| 9,0 | 0,02459 | 8,943 |
| 10,0 | 0,02459 | 8,943 |

Tabela 6.45 – Resultados para $\sigma_\rho$ e $\overline{T}$ - $\Delta = 0.03$

| Restrição $\overline{T}_{linit}$ (min) | *Fitness* $\sigma_\rho$ | $\overline{T}$ (min) |
|---|---|---|
| 7,8 | 0,0456 | 7,797 |
| 8,0 | 0,03436 | 7,99 |
| 8,2 | 0,02917 | 8,197 |
| 8,4 | 0,02697 | 8,4 |
| 8,5 | 0,0263 | 8,484 |
| 9,0 | 0,02451 | 8,945 |
| 10,0 | 0,02451 | 8,945 |

Note nas tabelas 6.44 e 6.45 que, por exemplo, se $\overline{T}_{linit}$ = 8,0 minutos (se este for o valor estabelecido para o nível de serviço oferecido aos usuários do sistema), a melhor solução para o balanceamento das cargas de trabalho dos servidores ($\sigma_\rho$) é 0,03436 e o tempo médio de resposta ($\overline{T}$) é 7,99 minutos. Observe também que, se $\overline{T}_{linit}$ < 7,796 (no caso de $\Delta = 0.05$) e $\overline{T}_{linit}$ < 7,797 (no caso de $\Delta = 0.03$), o problema não possui solução, dado que estes correspondem ao mínimo valor de $\overline{T}$. Note também, que ao estabelecermos $\overline{T}_{linit}$ > 8,943 (no caso de $\Delta = 0.05$) e $\overline{T}_{linit}$ > 8,945 (no caso de $\Delta = 0.03$), o problema corresponde a minimizar somente $\sigma_\rho$, dado que estes correspondem ao valor máximo de $\overline{T}$. Os gráficos das figuras 6.7 e 6.8 representam as curvas obtidas com os dados das tabelas 6.44 e 6.45, respectivamente.

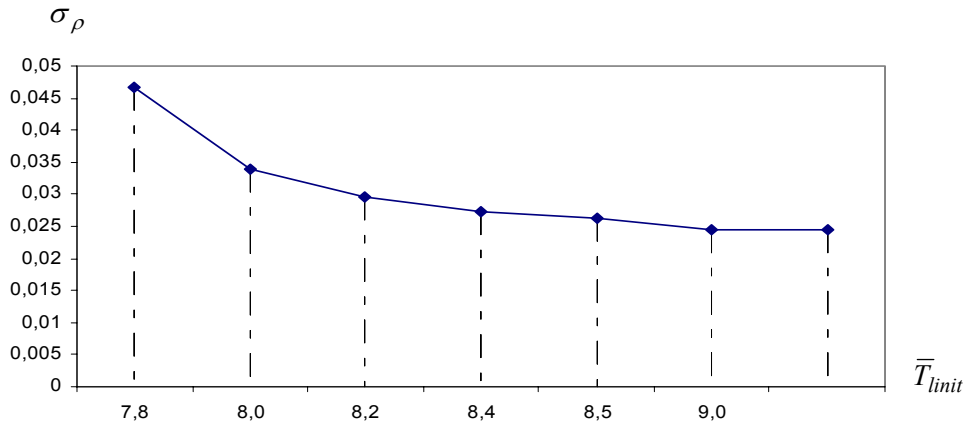


Figura 6.7 - Gráfico dos resultados de $\sigma_\rho$ x $\overline{T}_{linit}$ ($\Delta = 0.05$)

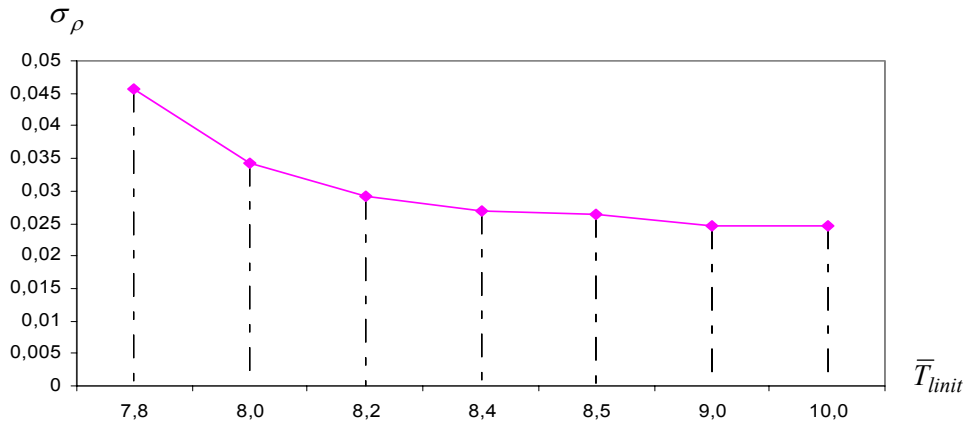


Figura 6.8 - Gráfico dos resultados de $\sigma_\rho$ x $\overline{T}_{linit}$ ($\Delta = 0.03$)

### 6.5.5 Resultados para instâncias com N >6 servidores:

Para verificar o desempenho da abordagem proposta para sistemas maiores que o sistema *Anjos do Asfalto* (N > 6), aplicamos este método para problemas testes cujos dados foram gerados aleatoriamente com base nos dados daquele sistema.

Para gerar cada problema teste, estabelecemos de forma aleatória a taxa de chegada em cada átomo *i* do sistema, utilizando a máxima e mínima taxa de chegada de chamadas do sistema *Anjos do Asfalto*, apresentados na tabela 6.1 onde: $\lambda_{\min} = 0{,}00008$ e $\lambda_{\max} = 0{,}00375$. Assim a taxa de chegada $\lambda_i$ em cada átomo *i* foi sorteada de uma distribuição uniforme no intervalo $(\lambda_{\min}, \lambda_{\max})$. Como discutido anteriormente, o número de átomos para sistemas com

configuração similar ao sistema *Anjos do Asfalto* deve ser $(2 \times N) - 2$ átomos, onde N é o número de servidores.

O mesmo procedimento foi utilizado para gerar os dados de entrada do processo de atendimento, e a taxa de atendimento $\mu_j$ de cada servidor *j* foi sorteada de uma distribuição uniforme no intervalo ($\mu_{\min}$ , $\mu_{\max}$), onde $\mu_{\min}$ = 0,0241 e $\mu_{\max}$ = 0,0101 (dados da tabela 6.2 do sistema *Anjos do Asfalto*). Para estabelecer o tamanho dos átomos do sistema, dividimos o trecho total da rodovia em átomos igualmente espaçados.

Desta forma, foram geradas instâncias para N= 6, N=8, N=10 e N=12 servidores. Inicialmente, utilizamos o modelo Hipercubo para calcular as medidas de performance da configuração original de cada instância, com foco nas três medidas objetivos descritas na seção 5.6 do capítulo 5: (i) desvio padrão das cargas de trabalho dos servidores; (ii) tempo médio de resposta para sistema; (iii) fração de chamadas atendidas em mais de 10 minutos.

Métodos de solução:

Como mencionado nas seções anteriores deste capítulo, na solução do sistema linear do modelo Hipercubo nas análises dos sistemas *Anjos do Asfalto* e *Centrovias* (com N = 6 e 5 servidores) utilizamos o método de Gauss-Jordan, dados que os tempos computacionais são razoáveis e o método é exato.

No entanto, ao resolver as instâncias com N =10 e N= 12, verificamos que o tempo computacional cresce de pouco tratável computacionalmente, dado que o número de equações é: $2^N$. Desta forma, a alternativa foi utilizar o método iterativo de Gauss-Siedel. Em CHIYOSHI et al (2000) e CHIYOSHI et al (2001) encontra-se uma descrição detalhada do uso deste método para solução do modelo Hipercubo exato. Nestes estudos, os autores ressaltam que, embora a matriz dos coeficientes do sistema de equações do modelo Hipercubo não satisfaça a condição de convergência do método, há duas considerações que podem justificar a conjectura de que este convirja: a normalização das incógnitas envolvidas e a experiência dos autores com dezenas de problemas testes.

No presente estudo, aplicamos o método para solução do modelo Hipercubo com os dados de entrada do sistema *Anjos do Asfalto* (apresentados na tabela 6.2) no início deste capítulo. Como descrito na seção 6.1, para este sistema, o valor de $\rho$ é 0,1890. Verificamos então que, gerando a matriz dos coeficientes do sistema de equações do modelo Hipercubo na seqüência $x_1 \rightarrow x_n$, o sistema não converge. Realizando, outros testes variando o valor de $\rho$, verificamos que este parâmetro influi na convergência do método, dado que para valores $\rho > 0{,}30$, há convergência. No entanto, em CHIYOSHI et al (2001), os autores explicam em detalhes porque a convergência do método depende da relação demanda/capacidade do problema em análise, e os efeitos de inversão na ordenação das variáveis na matriz para a convergência do método, ou seja, ao invés de $x_1 \rightarrow x_n$, sugere-se utilizar a seqüência $x_n \rightarrow x_1$. Realizando a inversão da matriz na solução do modelo Hipercubo aplicado ao presente estudo de caso, a convergência foi então obtida num número razoável de iterações, e o método foi aplicado às demais instâncias geradas. A tabela 6.46 apresenta os tempos computacionais para a solução destes problemas testes para cada método (exato e iterativo) utilizado.

Tabela 6.46– Tempo computacional para resolver o modelo Hipercubo (em segundos)

| Número de servidores N | Método Gauss-Jordan | Método Gauss-siedel |
|---|---|---|
| 6 | 0,02 | 0,02 |
| 8 | 0,16 | 0,06 |
| 10 | 10,66 | 1,28 |
| 12 | 579,79 | 14,38 |

Resultados da aplicação do modelo Hipercubo para problema teste:

De forma similar a análise descrita na seção 6.4 deste capítulo, aplicamos a abordagem GA/Hipercubo para cada problema teste, utilizando os parâmetros: tamanho da população (*Pop* = 100) para as instâncias com N =6, 8 e 10 servidores e (*Pop* = 50) para a instância com N= 12 servidores; número de gerações (*G* = 1000) para todas as instâncias; probabilidade de *crossover* (*pc* = 0.7) e probabilidade de mutação (*pm* = 0.06). Foram conduzidos experimentos utilizando individualmente as funções objetivo: $\overline{T}$ e $\sigma_\rho$, variando-se os intervalos: $\Delta$ = 0.01, 0.03 e 0.05. Os resultados obtidos são apresentados nas tabelas 6.47 a 6.50. No caso da configuração inicial, os valores de $y_j$ são iguais à 0,5. Para todas as instâncias foi utilizado o método de Gauss-Siedel e os experimentos para as instâncias 3 (N = 10 servidores) e 4 (N = 12 servidores) foram conduzidos em um microcomputador com

processador Pentium *IV* de 3.0 GHz (mais rápido que o Pentium de 2.0 GHz utilizado nos demais experimentos).

Instância 1 – N = 6 servidores: os resultados para a configuração original, e para a solução obtida pelos algoritmos GA/Hipercubo e enumerativo/Hipercubo, considerando os dois objetivos ($\overline{T}$ *e* $\sigma_\rho$), são apresentados na tabela 6.47. A melhor solução (configuração) obtida para o objetivo 1 é: $y_1$= 0.40; $y_2$=0.55; $y_3$=0. 40, $y_4$=0.65 e $y_5$=0.35 ($\Delta$ = 0.05), e para o objetivo 2 é: $y_1$= 0.35; $y_2$=0.44; $y_3$=0. 53, $y_4$=0.80 e $y_5$=0.20 ($\Delta$ = 0.03).

Tabela 6.47 – Medidas de desempenho para a configuração original e configurações ótimas

| Medidas | Original | Objetivo 1 | Melhora | Objetivo 2 | Melhora |
|---|---|---|---|---|---|
| $\overline{T}$ *(min)* | **6,735** | **6,197** | **7,98%** | 6,635 | 1,48% |
| $P_{tv>10}^{-}$ | **0,1806** | 0,1836 | -1,66% | 0,2519 | - 39,48% |
| $\sigma_\rho$ | **0,0576** | 0,0295 | 48,78% | **0,0024** | **95,83%** |

Note na tabela 6.47, que para o objetivo 1: a medida $\overline{T}$ ( a ser minimizada) é reduzida em 7,98% se comparada ao resultado da configuração inicial, enquanto $\sigma_\rho$ também melhora em 48,78% e $P_{tv>10}^{-}$ aumenta em apenas 1,66%. Observe que para o objetivo 1, ambas as medidas $\sigma_\rho$ e $\overline{T}$ são reduzidas em uma mesma configuração com relação à configuração original. No caso do objetivo 2: a medida objetivo ($\sigma_\rho$), que procuramos minimizar é reduzida de forma significativa em 95,83% com relação a configuração original, o tempo médio no sistema ($\overline{T}$) reduz-se apenas 1,48%, enquanto que a fração de chamadas atendidas em mais que 10 minutos ($P_{tv>10}^{-}$) aumenta 39,48%. O tempo computacional necessário para executar o algoritmo GA/Hipercubo foi em média de 185 segundos.

Instância 2 – N = 8 servidores: a tabela 6.48 apresenta os resultados para a configuração original, e para a solução obtida pelo GA/Hipercubo para os dois objetivos analisados. A melhor solução (configuração) obtida para o objetivo 1 é: $y_1$= 0.56; $y_2$=0.44; $y_3$=0. 53, $y_4$=0.53, $y_5$=0.44, $y_6$=0.41, $y_7$=0.59 ($\Delta$ = 0.03), e para o objetivo 2 é: $y_1$= 0.80; $y_2$=0.65; $y_3$=0.50, $y_4$=0.20, $y_5$=0.20, $y_6$=0.20, $y_7$=0.25 ($\Delta$ = 0.05)

Tabela 6.48 – Medidas de desempenho para a configuração original e configurações ótimas

| Medidas | Original | Objetivo 1 | Melhora | Objetivo 2 | Melhora |
|---|---|---|---|---|---|
| $\overline{T}$ *(min)* | **6,585** | **6,449** | **2,06%** | 7,392 | -12,26% |
| $P^{-}_{tv>10}$ | **0,1667** | 0,1730 | -3,78% | 0,3019 | -81,10 |
| $\sigma_\rho$ | **0,0554** | 0,0513 | 7,40% | **0,0358** | **35,38%** |

Os resultados da tabela 6.48 mostram que com relação ao objetivo 2, a medida minimizada $\overline{T}$ reduz-se em 2,06%, $\sigma_\rho$ reduz-se em 7,4% e $P^{-}_{tv>10}$ aumenta em 3,78 % com relação a configuração inicial. Como nos resultados para instancia 1 (N=6 servidores), ao minimizar $\overline{T}$ também reduzimos $\sigma_\rho$, enquanto que $P^{-}_{tv>10}$ aumenta. Ao minimizar $\sigma_\rho$ (objetivo 2), esta medida é reduzida em 35,38% com relação a configuração inicial, enquanto que $\overline{T}$ aumenta 12,26% e $P^{-}_{tv>10}$ aumenta significantemente em 81,10%. O tempo computacional necessário para executar o algoritmo GA/Hipercubo foi em média de 5.000 segundos (1,38 horas).

Instância 3 – N = 10 servidores: os resultados para esta instância são apresentados na tabela 6.49. A melhor solução (configuração) obtida para o objetivo 1 é: $y_1$= 0.32; $y_2$=0.41; $y_3$=0. 77, $y_4$=0.32, $y_5$=0.50, $y_6$=0.44, $y_7$=0.53, $y_8$=0.59, $y_9$=0.53 ($\Delta$ = 0.03), e para o objetivo 2 é: $y_1$= 0.41; $y_2$=0.32; $y_3$=0. 68, $y_4$=0.80, $y_5$=0.80, $y_6$=0.80, $y_7$=0.29, $y_8$=0.20, $y_9$=0.20 ($\Delta$ = 0.03).

Tabela 6.49 – Medidas de desempenho para a configuração original e configurações ótimas

| Medidas | Original | Objetivo 1 | Melhora | Objetivo 2 | Melhora |
|---|---|---|---|---|---|
| $\overline{T}$ *(min)* | **7,083** | **6,991** | **1,30%** | 7,873 | -11,15% |
| $P^{-}_{tv>10}$ | **0,2127** | 0,2396 | -12,65% | 0,3477 | -63,47% |
| $\sigma_\rho$ | **0,0699** | 0,0758 | -8,44% | **0,0454** | **35,05%** |

Note nos resultados da tabela 6.49, que quando a medida $\overline{T}$ é minimizada (objetivo 1), a redução obtida é de apenas 1,3%, e as medidas $\sigma_\rho$ e $\overline{T}$ apresentam piores valores, ou seja, aumentam em 8,44% e 12,65%, respectivamente. Ao minimizarmos $\sigma_\rho$ ocorre uma redução de 35,05% nesta medida quando comparada a solução original, e as medidas $\overline{T}$ e $P^{-}_{tv>10}$ aumentam em 11,15% e 63,47%, respectivamente. O tempo computacional necessário para executar o algoritmo GA/Hipercubo foi em média de 104.000 segundos (28,9 horas).

Instancia 4 – N = 12 servidores: os resultados para esta instância são apresentados na tabela 6.50. A melhor solução (configuração) obtida para o objetivo 1 é: $y_1$= 0.65, $y_2$=0.53; $y_3$=0.50, $y_4$=0.44, $y_5$=0.47, $y_6$=0.65, $y_7$=0.44, $y_8$=0.35, $y_9$=0.38, $y_{10}$=0.62, $y_{11}$=0.47 ($\Delta$ = 0.03), para o objetivo 2 é: $y_1$= 0.62, $y_2$=0.53; $y_3$=0.77, $y_4$=0.59, $y_5$=0.65, $y_6$=0.65, $y_7$=0.20, $y_8$=0.53, $y_9$=0.77, $y_{10}$=0.80, $y_{11}$=0.20 (($\Delta$ = 0.03).

Tabela 6.50 – Medidas de desempenho para a configuração original e configurações ótimas

| Medidas | Original | Objetivo 1 | Melhora | Objetivo 2 | Melhora |
|---|---|---|---|---|---|
| $\overline{T}$ *(min)* | **6,9097** | **6,8407** | **0,10%** | 7,341 | - 6,24% |
| $P_{\overline{tv}>10}$ | **0,1967** | 0,2192 | -11,44% | 0,3008 | -52,92% |
| $\sigma_\rho$ | **0,0725** | 0,0633 | 12,69% | **0,0443** | **38,90%** |

Na tabela 6.50 notamos que para o objetivo 1, a redução em $\overline{T}$ (medida que é minimizada) é pouco significativa, enquanto que $\sigma_\rho$ reduz-se em 12,69% e $P_{\overline{tv}>10}$ aumenta em 11,44% com relação a configuração inicial. Ao minimizar $\sigma_\rho$ (objetivo 2) esta medida é reduz-se em 38,9% com relação a configuração inicial, e as medidas $\overline{T}$ e $P_{\overline{tv}>10}$ aumentam em 6,24% e 52,92%, respectivamente. O tempo computacional necessário para executar o algoritmo foi em média de 150,15 horas (6,25 dias).

As análises do desempenho da abordagem GA/Hipercubo para as instâncias geradas aleatoriamente com N= 6,...,12 servidores mostram que, como na análise do sistema *Anjos do Asfalto*, as diferentes medidas de desempenho analisadas podem ser conflitantes quando apenas uma delas é considerada como função objetivo (*fitness*). Notamos nas tabelas 6.47 à 6.50 que, de forma similar aos *Anjos do Asfalto*, a medida $\overline{T}$ apresenta pequena melhora quando minimizada (objetivo 1) se comparada à melhora no valor de $\sigma_\rho$, quando esta é minimizada (objetivo 2). Além disso, ao aumentamos o número de servidores a redução obtida ao minimizarmos $\overline{T}$ decresce.

Notamos por meio desta análise que, mesmo utilizando um método iterativo para solução dos sistemas lineares do modelo Hipercubo, os tempos computacionais crescem significantemente (e de forma pouco tratável para sistemas com N > 10 servidores). No entanto, devido ao programa de concessão de trechos de rodovias a diversas concessionárias, são poucos os SAEs em rodovias brasileiras que utilizam N > 10 servidores Segundo dados da ABCR

(2005), apenas a concessionária Nova Dutra (na rodovia Presidente Dutra) possui N = 11 SAUs, sendo que esta corresponde a maior concessionária dos estados de São Paulo e Rio de Janeiro.

## 6.6 Algoritmo Enumerativo/Hipercubo para análise do SAE *Centrovias* :

Como o SAE *Centrovias* possui apenas N = 5 servidores, integramos o modelo Hipercubo múltiplo despacho em um algoritmo de enumeração completa, de forma a determinar uma configuração ótima (ou perto da ótima) para o sistema. O algoritmo foi implementado de forma similar ao algoritmo enumerativo/Hipercubo descrito na seção 6.5.2. Detalhes deste experimento também podem ser encontrados em IANNONI et al (2005). No entanto, para sistemas maiores, o método GA/Hipercubo proposto no capítulo 5 seria o mais indicado, dado as limitações computacionais do procedimento de enumeração.

O procedimento que determina as possíveis configurações do sistema é similar ao descrito na seção 5.6 do capítulo 5 para o algoritmo GA/Hipercubo. Desta forma, cada configuração do sistema é representada por um vetor $y = (y_1,y_2,.....y_{N-1})$, onde $y_j$ é a fração da distância entre dois servidores adjacentes $j$ e $j$+1, que corresponde a uma das áreas preferenciais do servidor $j$. Conforme o algoritmo descrito no capítulo 5, consideramos $0,2 < y_j < 0,8$, de forma a limitar a área preferencial de cada servidor, respectivamente, em 20 e 80 por cento da distância entre dois servidores adjacentes, e adicionamos a 0.2 (limite inferior) o incremento $k\Delta$, onde $\Delta$ é fixo e $k = 0,1,...,M= (0,8 – 0,2)/ \Delta$. Utilizando $\Delta = 0,05$ e $\Delta = 0,03$, como na seção 6.3.2, temos $(M+1)^{N-1}$ combinações possíveis ($13^4$ e $21^4$, respectivamente).

Para gerar os dados de entrada de cada configuração (taxa de chegada nos átomos do sistema de forma a preservar a distribuição das taxas de chegada ao longo da rodovia e tempo médio de viagem servidor - átomo), utilizamos o mesmo procedimento descrito pela expressão 5.3 e figura 5.1 na seção 5.6 do capítulo 5. Conforme a seção 6.2, ao implementarmos o modelo Hipercubo para analisar o sistema *Centrovias*, utilizamos os dados disponíveis na amostra para calcular os tempos médios de viagem de cada servidor a cada átomo, considerando a distribuição da localização das chamadas ao longo da rodovia. No entanto, como discutido na seção 5.6 (capítulo 5), o procedimento que varia o tamanho do átomo resulta em variações nos centróides de cada átomo, e assim os tempos de viagem devem ser recalculados a cada nova

configuração. Desta forma, ao implementar o algoritmo enumerativo/Hipercubo, recalculamos a matriz dos tempos de viagem servidor – átomo da configuração inicial e das configurações alternativas, de acordo com a distância da base do servidor ao centróide do átomo (como proposto em LARSON & ODONI, 1981).

De forma similar a seção 6.5, conduzimos diferentes experimentos utilizando 3 diferentes medidas de desempenho do sistema para determinar qual a melhor dentre as configurações geradas. Assim, estas três medidas de desempenho são (i) tempo médio de viagem no sistema (min) - $\overline{T}$, (ii) fração de chamadas atendidas em tempo superior à 10 min - $P_{\bar{tv}>10}$ e (iii) desvio padrão das cargas de trabalho - $\sigma_\rho$. As expressões que determinam $\sigma_\rho$ e $P_{\bar{tv}>10}$ foram apresentadas, nos capítulos 5 (expressão 5.6) e 4 (expressão 4.56), respectivamente. Para determinar a medida do tempo médio de viagem no sistema é preciso definir uma medida que agrega as medidas de tempo de viagem tipo 1 e tipo 2 apresentadas na seção 4.2 do capítulo 4. Assim utilizamos:

$$\overline{T} = \sum_{i=1}^{N_A}\left( \sum_{j=1}^{N}\left[ f_{ji}^{'[1]} + f_{ji}^{'[2]} \right].t_{ji} + \sum_{j=1}^{N-1}\sum_{k=j+1}^{N} f_{(j,k)i}^{'[2]} \min(t_{ji}, t_{ki}) \right),$$ onde $f_{ji}^{'[1]}, f_{ji}^{'[2]}, f_{(j,k)i}^{'[2]}$ e $t_{ji}$ foram definidos na seção 4.2 do capítulo 4 (expressões 4.34. a 4.36).

A configuração inicial do sistema em termos do tamanho dos átomos no sistema pode ser representada pelo vetor *y* da seguinte forma: $y_1$ = 0,61; $y_2$ =0,28; $y_3$ =0,47 e $y_4$ = 0,61. Os resultados obtidos para as três medidas utilizadas como medidas objetivo na configuração inicial são: $\overline{T}$ = 5,958 min, $P_{\bar{tv}>10}$ = 0,3119 e $\sigma_\rho$ = 0,0173.

Realizamos experimentos com Δ = 0,05 e Δ = 0,03, e para os três objetivos as melhores soluções foram encontradas com Δ = 0,03. As melhores configurações para cada objetivo são:

(i) objetivo 1 - Tempo médio de viagem no sistema ($\overline{T}$)

$y_1$ = 0.23; $y_2$ =0.32; $y_3$ =0.62 e $y_4$ = 0.53

(ii) objetivo 2 - fração de chamadas atendidas em tempo superior à 10 min ($P_{\bar{tv}>10}$).

$y_1$ = 0.44; $y_2$ =0.20; $y_3$ =0.74 e $y_4$ = 0.50

(iii) objetivo 3 - desvio padrão das cargas de trabalho ($\sigma_\rho$) :

$y_1$ = 0.62; $y_2$ =0.20; $y_3$ =0.59 e $y_4$ = 0.53

A tabela 6.51 apresenta a melhor solução (em negrito) em termos de cada objetivo e a porcentagem de melhora com relação a configuração original.

Tabela 6.51 – Medidas de desempenho da melhor solução

| Medidas | Original | Objetivo 1 | Melhora | Objetivo 2 | Melhora | Objetivo 3 | Melhora |
|---|---|---|---|---|---|---|---|
| $\overline{T}$ *(min)* | **5,9584** | **3,4054** | **42,85%** | 4,8047 | 19,36% | 5,9654 | - 0,12% |
| $P_{tv>10}^{-}$ | **0,3119** | 0,3472 | -11,32% | **0,2779** | **10,90%** | 0,2970 | 4,78% |
| $\sigma_\rho$ | **0,01733** | 0,0207 | -19,73% | 0,01732 | 0,0% | **0,0163** | **6,00%** |

Note na tabela 6.51 que no primeiro experimento, $\overline{T}$ (que procuramos minimizar) apresenta significativa redução de 42,85%, com relação a configuração original. No entanto, $\sigma_\rho$ e $P_{tv>10}^{-}$ aumentam em 19,73% e 11,32%, respectivamente. No segundo experimento, a medida a ser minimizada $P_{tv>10}^{-}$ reduz-se em 10,9%, enquanto que $\overline{T}$ também é reduzido em 19,36% e $\sigma_\rho$ não sofre alterações significativas (redução de apenas 0,5%). No último experimento, $\sigma_\rho$ (medida que é minimizada) reduz-se em 6,0% com relação a configuração inicial, enquanto que $\overline{T}$ aumenta em 0,12%. Estes resultados mostram que, como na análise realizada na seção 6.5 para o sistema *Anjos do Asfalto*, ao consideramos os diferentes objetivos separadamente, as medidas avaliadas podem ser conflitantes. Por exemplo, a melhor solução em termos de $\overline{T}$ resulta em piores valores para $\sigma_\rho$ e $P_{tv>10}^{-}$.

# 7. Conclusões e Perspectivas

## 7.1 Conclusões

Este estudo mostra como o modelo Hipercubo de filas espacialmente distribuídas pode ser modificado e aplicado para análise dos sistemas de atendimento médico emergencial em rodovias, considerando a aleatoriedade envolvida na operação destes sistemas e suas particularidades com relação a política de despacho de ambulâncias. Além disso, considerando que o modelo é essencialmente descritivo, este estudo também propõe uma abordagem que integra o modelo Hipercubo em um algoritmo genético para analisar os SAEs em rodovias. Por meio desta abordagem é possível prescrever qual a configuração ótima (ou perto da ótima) do SAE analisado, em termos das principais medidas de desempenho deste sistema. Um dos aspectos inovadores desta abordagem é tratar do dimensionamento das áreas de cobertura de cada servidor, tal que otimize a configuração e operação do sistema em termos das medidas de desempenho mais relevantes do ponto de vista dos usuários e dos operadores do sistema. Por exemplo, minimizar o desbalanceamento das cargas de trabalho entre os servidores ou/e minimizar o tempo médio de resposta aos usuários do sistema.

Para aplicação dos métodos propostos, utilizamos dois estudos de caso: O primeiro é o SAE *Anjos do Asfalto* na rodovia Presidente Dutra, que foi inicialmente estudado por MENDONÇA & MORABITO (2000, 2001). Para analisar este SAE, utilizamos os dados da pesquisa de campo e as informações daquele estudo. Este sistema é caracterizado por uma política de único despacho de ambulância com *backup* parcial, e foi utilizado como base para a implementação inicial da presente abordagem, que combina um algoritmo genético com o modelo Hipercubo.

Na implementação da abordagem GA/ Hipercubo realizamos três experimentos com três diferentes funções de aptidão (*fitness),* representando três objetivos: (i) desvio padrão das cargas de trabalho dos servidores; (ii) tempo médio de resposta aos usuários do sistema; (iii) fração de chamadas atendidas com tempo de viagem maior que 10 minutos. Além disso, devido ao moderado tamanho do sistema analisado (N = 6 servidores), também definimos um algoritmo de enumeração completa para prescrever a solução ótima com precisão de $\Delta$ (parâmetro utilizado no procedimento de variação do tamanho dos átomos e geração de diferentes configurações do sistema). Os resultados obtidos mostraram que abordagem

GA/Hipercubo é efetiva para encontrar a solução ótima prescrita pelo algoritmo enumerativo/Hipercubo. A abordagem GA/Hipercubo utilizando diferentes funções *fitness* mostrou que há conflitos entre as medidas de desempenho, especialmente entre o balanceamento das cargas de trabalho e a minimização do tempo médio de resposta aos usuários do sistema. Sendo assim, tratamos o *trade-off* entre as medidas de desempenho utilizando um simples método de otimização bi-objetivo (método $\varepsilon$ -restrito) que escolhe um dos objetivos como função *fitness* do GA/Hipercubo, e as soluções geradas passam a ser sujeitas a restrição relacionada ao outro objetivo.

Ao realizar experimentos para sistemas com maior número de servidores, verificamos que a abordagem GA/Hipercubo torna-se inviável computacionalmente a medida que $N$ cresce, mesmo utilizando métodos iterativos na solução dos sistemas lineares do modelo Hipercubo. Isso sugere que outras alternativas devem ser testadas para aprimorar esta abordagem tais como: utilizar o modelo Hipercubo aproximado (LARSON, 1975 e JARVIS, 1985) e/ou outras meta-heurísticas que avaliam apenas uma solução a cada iteração, e não uma população de soluções como o algoritmo genético. Convém salientar, que a extensão do modelo Hipercubo aproximado para o caso de SAEs em rodovias não é trivial, particularmente por causa do *backup* parcial, múltiplo despacho de ambulâncias, entre outras particularidades destes sistemas.

O segundo estudo de caso é o SAE de rodovias do interior do Estado de São Paulo, sob administração da concessionária *Centrovias*. Entre as principais particularidades da operação deste sistema estão: política de múltiplo despacho de ambulâncias e *backup* parcial, servidores ocupados atendendo chamadas que não são de emergência (chamadas atendidas na base do servidor) e servidores diferenciados (p.e, carro médico e veículo resgate). Os resultados obtidos pelo modelo Hipercubo, modificado para tratar tais particularidades, mostraram que o método é eficaz/efetivo para avaliar as principais medidas de desempenho do sistema. Estes resultados foram validados por meio da análise de amostras e resultados da simulação discreta do sistema.

Com base na configuração do sistema *Centrovias* (de tamanho moderado, com apenas N = 5 servidores), também propomos um método enumerativo integrado ao modelo Hipercubo múltiplo despacho, de forma a determinar o dimensionamento das áreas de operação das ambulâncias do sistema. Os resultados mostraram que as medidas de desempenho do sistema

podem ser melhoradas de forma significativa com a simples variação do tamanho dos átomos do sistema, sem que sejam necessários investimentos adicionais ao sistema. Como no caso anterior, este método também pode ser utilizado para avaliar o *trade-off* entre as medidas de desempenho conflitantes do sistema. Convém ressaltar que para sistemas maiores, uma abordagem que utilize meta-heurísticas seria o mais indicado para otimizar a configuração deste SAE, por exemplo, a própria abordagem GA/Hipercubo adaptada para analisar o sistema *Anjos do Asfalto*.

**7.2 Perspectivas**:

Entre as perspectivas deste estudo está a aplicação dos modelos e métodos propostos para outros SAEs em rodovias, com características similares aos SAEs analisados no presente estudo. Em particular, sistemas com múltiplo despacho de ambulâncias idênticas e/ou diferenciadas. Como mencionado anteriormente, para sistemas maiores (p.e, N >> 10 servidores), a abordagem que integra o modelo Hipercubo em um método de otimização poderia ser aprimorada com a utilização, por exemplo, de outras meta-heurísticas tais como: Busca Tabu, *Simulated Annealing*, GRASP, entre outras.

Para otimizar a configuração destes sistemas avaliando o *trade-off* entre as diferentes medidas objetivo, outros métodos de otimização multiojetivo podem ser inseridos nos procedimentos de busca da meta-heurística utilizada, tais como os métodos baseados na geração de soluções eficientes de Pareto para algoritmos genéticos e Busca Tabu multiobjetivo.

Convém ressaltar que as adaptações do modelo Hipercubo propostas para análise dos SAEs em rodovias podem ser diretamente aplicados em abordagens de otimização para localização dos servidores, por exemplo, nos métodos propostos em BATTA et al (1989), CHIYOSHI et al. (2003), SAYDAM & AYTUG (2003) e GALVÃO et al. (2005).

Uma pesquisa futura interessante da abordagem GA/Hipercubo desenvolvida é sua extensão para otimizar a configuração e operação dos SAEs em rodovias por meio das decisões combinadas de dimensionamento dos átomos do sistema, e de localização dos *N* servidores do sistema. Assim, a questão é como tratar das duas decisões (localização e dimensionamento dos átomos) em uma mesma abordagem.

## Referências:


ABCR (2005) Associação Brasileira das Concessionárias de Rodovias. www.abcr.org. (acesso 02/2005).

ALBINO J.C.C. (1994) Quantificação e locação de unidades móveis de atendimento de emergência e interrupções em redes de distribuição de energia elétrica: aplicação do Modelo Hipercubo. *Dissertação* (Mestrado em Engenharia de Produção) - Departamento de Engenharia de Produção, Florianópolis : UFSC.

AYTUG H., SAYDAM C. (2002) Solving large-scale maximum expected covering location problems by genetic algorithms: A comparative study. *European Journal of Operations Research* 141, p. 480-494.

ARROYO, J.E.C. (2002) Heurísticas e metaheurísticas para otimização combinatória multiobjetivo. Universidade Estadual de Campinas. *Tese* (doutorado em Engenharia Elétrica) - Faculdade de Engenharia Elétrica e Computação.

ARTESP (2005) Agência Reguladora de Serviços Públicos Delegados de Transporte do Estado de São Paulo. www.artesp.sp.gov.br (acesso 02/2005).

BANKS J. (1998) *Handbook of Simulation*. John Wiley & Sons, Atlanta.

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

BEASLEY J.E., CHU P.C. (1996) A genetic algorithm for the set covering problem. *European Journal of Operations Research* 94, p. 392-404.

BEASLEY J.E. (2000) A population heuristic for constrained two-dimensional non-guillotine cutting. Disponível em http://mscmga.ms.ic.ac.uk/jeb/jeb.html.(acesso 04/2004).

BEASLEY J.E. (2004) A population heuristic for constrained two-dimensional non-guillotine cutting. *European Journal of Operational research* 156, p. 601-627

BELL C., ALLEN D. (1969) Optimal planning of an emergency ambulance service. *Socio - Economic Planning Sciences* 3 (2), p. 95 - 101.

BODILY S. (1978) Police sector design incorporating preferences of interest groups for equality and efficiency. *Management Science* 24(12), p.1301 – 1313.

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

BROTCORNE L., LAPORTE G., SEMET F. (2003). Ambulance location and relocation models. *European Journal of Operational research* 147, p.451-63.

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

CENTROVIAS (2004) – www.centrovias.com.br, (acesso 12/2004).

CHAIKEN J. (1971) The number of emergency units busy at alarms which require multiple servers. The Rand Corporation, Santa Monica, CA.

CHAIKEN J., LARSON R. (1972) Methods for allocating urban emergency units. *Management Science* 19(4), p.110 – 130.

CHAIKEN J., DORMONT P. (1978a) A patrol car allocation model: Background. *Management Science* 24(12), p.1280 – 1290.

CHAIKEN J., DORMONT P. (1978b) A patrol car allocation model: Capabilities and algorithms. *Management Science* 24(12), p.1291 – 1300.

CHELST K.R. (1975) Implementing the Hypercube Queuing Model in the New Haven Department of Police Services: A case study in Technology Transfer. *The Rand Corp*. Santa Monica. C.A.

CHELST K.R. (1978) An algorithm for deploying a crime directed patrol force. *Management Science* 24(12), p.1314 – 1327.

CHELST K.R., JARVIS J.P. (1979) Estimating the probability distribution of travel times for urban emergency service systems. *Operations Research* 27(1), p. 199-204.

CHELST K.R. (1981) Deployment of one-vs-two-officer patrol units: A comparison of travel times. *Management Science* 27(2), p. 213 – 230.

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

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

CHIYOSHI F., GALVÃO, R. D. (2000) A statistical analysis of simulated annealing applied to the p-median problem. *Annals of Operations Research* 96, p. 61-74.

CHIYOSHI F., GALVÃO R. D., MORABITO R.. (2001) Modelo Hipercubo: análise e resultados para o caso de servidores não-homogêneos. *Pesquisa Operacional* 21(2), p.199-218.

CHIYOSHI F., GALVÃO R. D., MORABITO R. (2003) A note on solutions to the maximal expected covering location problem. *Computers and Operations Research* 30 (1), p. 87-96.

CHRISTOFIDES N., VIOLA P. (1971) The optimum location of multi-centers on a graph. *Operation Research Quarterly* 22, p.145.

CHURCH R. L., REVELLE, C. (1974) The maximal covering location problem. *Papers Reg. Sci. Assoc*. 32, p.101-118.

COHON, J.L (1978) *Multiobjective programming & planning*. New York: Academic Press.

COSTA NETO P.L.O. (1977) *Estatística*. Edgar Blucher. São Paulo.

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

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

DASKIN M.S. (1995) *Network and Discrete Location*. Wiley & Sons. New York.

EATON D. J., DASKIN M. S., SIMMONS D., BULLOCH B., JANSMA G. (1985) Determining emergency medical service vehicle deployment in Austin, Texas. *Interfaces* 15(1), p. 96-108.

FITZSIMMONS J. A. (1973) A methodology for emergency ambulance deployment. *Management Science* 19(6), p. 627-636.

GALVÃO R.D., CHIYOSHI F., ESPEJO L.G.A., RIVAS M.P.A. (2003) Solução do problema de localização de máxima disponibilidade utilizando o modelo hipercubo. *Pesquisa Operacional*, *SOBRAPO* 23 (1), p. 61-78.

GALVÃO R.D., CHIYOSHI F., MORABITO R. (2005) Towards unified formulations and extensions of two classical probabilistic location models. *Computers & Operations Research* 32, 15-33.

GLOVER F., KELLY; J.P., LAGUNA M. (1995) Genetic algorithms and tabu search: hybrids for optimization. *Computers and Operations Research* 1, p. 111-134.

GOLDBERG J., DIETRICH J., CHEN J., MITWASI M., VALENZUELA T., CRISS E., (1990). A simulation model for evaluating a set of emergency vehicle base locations: Development validation and usage. *Socio-Economics Planning Science* 24, 125-141.

GOLDBERG D. E. (1989) *Genetic Algorithms in Search, Optimization and Machine Learning*. New York: Addison Wesley.

GONÇALVES M.B., NOVAES A.G., ALBINO J.C.C. (1994) Modelos para localização de serviços emergenciais em rodovias. In: Simpósio Brasileiro de Pesquisa Operacional 26, Florianópolis, SC, 1994. *Anais*. Florianópolis, p.591-596.

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

GREEN L. (1984) A multiple dispatch queuing model of police patrol operations. *Management Science*, 30(6), p. 653 – 664.

GREEN L., KOLESAR P. (1984a) A comparison of the multiple dispatch and M/M/C priority queuing models of police patrol. *Management Science*, 30(6), p. 665 – 670.

GREEN L., KOLESAR P. (1984b) The feasibility of one-officer patrol in New York City. *Management Science* 30(8), 964 – 981.

GRENDREAU M., LAPORTE G., SEMET F. (1997) Solving an ambulance location model by Tabu Search. *Location Science* 5, p. 75-88.

GRENDREAU M., LAPORTE G., SEMET F. (2001) A dynamic model and parallel Tabu search heuristic for real-time ambulance relocation. *Parallel Computing* 27, p.1641-1653.

HALPERN J. (1977) Accuracy of estimates for the performance criteria in certain emergency service queuing systems. *Transportation Science* 11(3). p. 223-242.

HERTZ A., KOBLER D. (2000) A framework for the description of evolutionary algorithms. *European Journal of Operational Research* 126, p.1-12.

HOLLAND J.H. (1975) *Adaptation in Natural and Artificial Systems*, MIT Press, Cambridge, MA.

IANNONI, A. P., MORABITO R. (2002). Análise do Sistema Logístico de Recepção de Cana de Açúcar: Um Estudo de Caso Utilizando Simulação Discreta. *Gestão & Produção* 9(2), p. 107-128.

IANNONI A.P, MORABITO R. (2004) A discrete simulation analysis of a Logistics Supply System. Aceito para publicação em *Transportation Research Part E*.

IANNONI A.P, MORABITO R., SAYDAM C. (2005) Analyzing the configuration and operation of emergency medical systems on highways using the hypercube model (submetido para publicação)

IGNALL E., KOLESAR A. SWERSEY A., WALKER W., BLUM G., CARTER G. (1975) Improving the deployment of the New York City fire companies. *Interfaces* 2(2), p. 48-61.

IGNALL E., KOLESAR A,. WALKER W. (1978) Using simulation to develop and validate analytic models: Some case studies. *Operations Research* 26, p. 237-253.

IGNALL E., CARTER G., RIDER K. (1982) An algorithm for the initial dispatch of fire companies. *Management Science* 28(4), p.366-378.

JARAMILLO J. H.; BHADURY J., BATTA R. (2002) On the use of Genetic algorithms to solve location problems, *Computers & Operations Research* 29, p. 761-779.

JARVIS J.P. (1985) Approximating the equilibrium behavior of multi-server loss systems. *Management Science* 31, p. 235 - 239.

JASKIEWICZ A. (2002) Genetic local search for multi-objective combinatorial optimization. *European Journal of Operational Research* 137, p.50-71.

KARASAKAL O., KARASAKAL E. K. (2004) A Maximal Covering Location Model in the presence of partial coverage. *Computers & Operations Research* 31(9), p.1515-1526.

KOLESAR P., BLUM E. (1973) Square root laws for fire engines response distances. *Management Science* 19(12), p. 1368 - 1378.

KOLESAR P., SWERSEY A. J. (1986) The deployment of urban emergency units: a survey. In: *Management science and delivery of urban services*, eds: SWERSEY A., IGNALL E. TIMS Studies in the Management Science 22, p. 87-119. Elsevier. North-Holland.

KELTON W. D., SADOWSKI R.P., SADOWSKI D. A (2002) *Simulation with Arena.* 2.ed, McGrawHill. New York.

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

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

LARSON R.C., ODONI A.R. (1981) *Urban operations research*. Prentice Hall. New Jersey.

LARSON R., MCKNEW M.A. (1982) Police patrol-initiated activities within a systems queuing model. *Management Science* 28(7), p. 759 – 774.

LARSON R.C. (2004) OR models for homeland security. *OR/MS Today* 31, 22-29.

LAW A. M., KELTON W. D. (1991) *Simulation Modeling and Analysis.* 2.ed. McGraw-Hill. New York.

LOUVEAUX F. (1993) Stochastic location analysis. *Location Science* 1, p.127-154.

MAGALHÃES M. N., LIMA A.P. (2001) *Noções de Probabilidade e Estatística*. 3.ed. Instituto de Matemática e Estatística IME-USP. São Paulo.

MARIANOV V., REVELLE C. (1996) The queuing maximal availability location problem: A model for the sitting of emergency vehicles. *European Journal of Operations Research*, p.110-120.

MENDONÇA F.C. (1999) Aplicação do modelo hipercubo, baseado em teoria de filas, para análise de um sistema médico-emergencial em rodovia. São Carlos: UFSCar, 1999. 112p. *Dissertação* (mestrado em Engenharia de Produção) - Departamento de Engenharia de Produção.

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

MENDONÇA F.C., MORABITO R. (2001) Analysing emergency service ambulance deployment on a Brazilian highway using the hypercube model. *Journal of the Operational Research Society* 52, p.261- 268.

MICHALEWICZ Z. (1996) *Genetic Algorithms + Data Structures = Evolution Programs*, 3rd ed., Springer-Verlag. Berlin.

MICHALEWICZ Z., DASGUPTA D., LE RICHIE R., SCHOENAUER M. (1996) Genetic algorithms for constrained engineering problems. *Computers and Industrial Engineering Journal* 30 (4), p. 851-870.

MICROCAL ORIGIN (1997). Microcal Softawre Inc.versão 5.0. www.microcal.com

MORABITO R., CHIYOSHI F., GALVÃO R.D. (2004) Homogeneous servers in Emergency Medical Systems: Practical applications using the hypercube queuing model. (submetido para publicação).

MINIEKA E. (1970) The M-centre problem. *SIAM Review* 12.

OWEN S., DASKIN M.S. (1998) Strategic facility location: a review. *European Journal of Operational Research* 3(2), p. 423-447.

PC - PALISADE CORPORATION. *Best Fit for Windows*: *Help*. 2.ed. New York, 1996.

PEGDEN C. D., SHANNON R. E., SADOWSKI R. P. (1995) *Introduction to Simulation Using SIMAN*. 2.ed. McGraw-Hill, New York.

PILEGGI, G.C.F (2002) Abordagens para otimização integrada dos problemas de geração e sequenciamento de padrões de corte. *Tese* (doutorado em Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação – USP.

REEVES C. R. (1997) Genetic Algorithm for the Operations Researcher. *ORMS Journal on Computing* 3, p. 231-250.

REVELLE C., MARKS D., LIEBMAN J.C. (1970) An analysis of private and public sector location models. *Management Science* 16(11), p. 692 – 707.

REVELLE C., HOGAN K. (1989) The maximum availability location problem. *Transportation Science* 23(3), p. 192-200.

REVELLE C. (1989) Review, extension and prediction in emergency service sitting models, *European Journal of Operations Research* 40, p.58-69.

RIDER K. (1976) A paramedic model for the allocation of fire companies in New York City. *Management Science* 23(2), p.146 – 158.

RIVAS M. P. A. (2002) Problemas Probabilísticos de Localização com restrições de cobertura utilizando o Modelo Hipercubo. COPPE/UFRJ, *Tese* (Doutorado em Engenharia de Produção) - Departamento de Engenharia de Produção.

SCHILLING D., ELZINGA D.J., COHAN J., CHURCH R., RVELLE C. (1979) The team/fleet models for simultaneous facility and equipment siting. *Transportation Science*, 13(2), p.163-175.

SAYDAM C., MCKNEW M. (1985) A separable programming approach to expected coverage: An application to ambulance location. *Decisions Sciences* 16, p.381-397.

SAYDAN C., REPEDE J., BURWELL T. (1994) Accurate estimation of expected coverage: a comparative study. *Socio -Economic Planning Sciences* 28 (2). p.113 - 120.

SAYDAM C., AYTUG H. (2003) Accurate estimation of expected coverage: revisited. *Socio -Economic Planning Sciences* 37, p. 69-80.

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

SAVAS E. (1969) Simulation and cost-effectiveness analysis of New York's emergency ambulance service. *Management Science* 15(12), B608 –B627.

SECTRAN (2005) Secretaria de Estado dos Transportes. www.sectran.sp.gov.br (acesso 05/2005).

STEUR, R.E. (1986) *Multiple criteria optimization: theory, computation & application*. New York: John Wiley.

SMC - Systems Modeling Corporation. (1994) *ARENA User's Guide*, Sewickley.

SWERSEY A.J. (1982) A Markovian decision model for deciding how many fire companies to dispatch. *Management Science* 28(4), p.352-365.

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

TAYLOR I.D., TEMPLETON J.G. (1980) Waiting time in a multi-server cutoff-priority queue, and its application to an urban ambulance service. *Operations Research* 28(5), p. 199-204.

TAKEDA R.(2000) Uma contribuição para avaliar o desempenho de sistemas de transporte emergencial de saúde. Escola de Engenharia de São Carlos – USP. *Tese* (doutorado em Engenharia de Transportes) - Departamento de Transportes.

TAKEDA R. A., WIDMER, J. A., MORABITO, R. (2000) Uma proposta alternativa para avaliação do desempenho de sistemas de transporte emergencial de saúde brasileiros. *Transportes* 9(2), p.9-27.

TAKEDA R. A., WIDMER, J. A., MORABITO, R. (2004) Analysis of ambulance decentralization in an urban medical emergency service using the hypercube queuing model. Aceito para publicação em *Computers & Operations Research*.

TAVAKOLI A., LIGHTNER C. (2004) Implementing a mathematical model for locating EMS in Fayetteville, NC. *Computers & Operations Research* 31, p.1549-1563.

TEIZ M., BART P. (1968) Heuristic methods for estimating the generalized vertex median of a weighted graph. *Operations Research* 16(5), p. 901-1092.

TOREGAS C., SWAIN C., REVELLE C., BERGMAN L. (1971) The location of emergency service facilities. *Operations Research* 19(6), p.1363-1373.

ZAKI A.S., CHENG H.K. (1997) A Simulation Model for the Analysis and Management of An Emergency Service System. *Socio-Economic Planning Sciences* 31, p.173-189.

# Anexos

**Formulação de alguns modelos de localização probabilística:**

## A1. Problema de Localização de Máxima Cobertura Esperada (MEXCLP – *Maximum Expected Covering Location Problem*) :

Como o modelo assume que os servidores são independentes, $\rho$ é fração de ocupação conhecida e igual para todos os servidores. A probabilidade de que um certo nó de demanda possa ser atendido por pelo menos um servidor disponível dado que há $j$ servidores oferecendo cobertura (ou seja, localizados a menos da distância crítica de S deste nó), é dada por: Pr[um ou mais servidores disponíveis a menos de $S$ ] = (1 – Pr[não há servidores disponíveis entre os $j$ servidores que cobrem o nó]) = $1-\rho^j$ .

Se definirmos as variáveis: $h_i$ como o número de chamadas no nó $i$ e $H_{i,j}$ como a demanda no nó $i$, coberta por um servidor disponível dado que $j$ servidores cobrem o nó $i$, então temos que:

$$H_{i,j} = \begin{cases} h_i & \text{com probabilidade 1 - } \rho^j \\ 0 & \text{com probabilidade } \rho^j \end{cases} \qquad \text{(A.1)}$$

E o valor esperado de $H_{i,j}$ é $E(H_{i,j}) = h_i(1-\rho^j)$ $\forall$ k,m, sendo que E($H_{i,j}$) representa a cobertura esperada no nó $i$, que é incrementada de $\Delta E(H_{i.j})$ quando o número de servidores que cobrem $i$ aumenta de $j-1$ para $j$. O incremento é expressado por:

$$\Delta E(H_{i.j}) = E(H_{i,j}) - E(H_{i,j-1}) = h_i \rho^{j-1}(1-\rho) \qquad \text{(A.2)}$$

Considerando estas definições, o MEXCLP é formulado da seguinte forma:

$$\text{Maximizar } \sum_{i \in I} \sum_{j=1}^{p} (1-\rho)\rho^{j-1} h_i x_{ij} \qquad \text{(A.3)}$$

$$\text{s.a} \sum_{j=1}^{p} x_{ij} \leq \sum_{k \in J} a_{ik} y_k \ , \ \forall i \in I \qquad \text{(A.4)}$$

$$\sum_{k \in J} y_k \leq p \, , \qquad \text{(A.5)}$$

$$y_k = 0,1,..,p \quad \forall k \in J \qquad \text{(A.6)}$$

$$x_{ij} = \{0,1\} \qquad \forall i, j \qquad \text{(A.7)}$$

onde,

$p$ : número de servidores que devem ser localizados;

$h_i$ : quantidade de demanda no nó $i$ ;

$x_{ij}$ = 1 se o nó $i$ é coberto por pelo menos $j$ servidores, e $x_{ij}$ = 0 se caso contrário;

$a_{ik}$ = 1 se o nó $i$ é coberto pelo servidor localizado no nó $k$ ;

$y_k$ : número de servidores localizados no nó $k$ ;

$I$ = conjunto de nós de demanda;

$J$ = conjunto de locais em potencial onde os servidores podem ser localizados.

A função objetivo busca maximizar a cobertura esperada sobre todos os nós de demanda, dado que há $p$ servidores para serem localizados. As restrições (A.4) estabelecem que o nó de demanda *i* pode ser coberto no máximo por $p$ servidores. As restrições (A.5) estabelecem que o número de servidores localizados pode ser no máximo $p$, sendo que mais de um servidor pode ser localizado em um mesmo nó. As restrições (A.6) expressam o caráter inteiro do número de servidores localizados em cada nó da rede. As restrições (A.7) mostram o caráter binário das variáveis de decisão.

## A2. Problema de Localização de Máxima Disponibilidade I (MALP I – *Maximum Availability Location Problem*)

A fração de ocupação, idêntica para todos os servidores é obtida considerando:

$\rho =$ tempo médio total de atendimento por dia/ tempo total de funcionamento do sistema por dia.

A fórmula desenvolvida por DASKIN (1983) para calcular esta taxa é:

$$\rho = \frac{\bar{t}.\sum_{i \in I} h_i}{24.p} \qquad \text{(A.8)}$$

onde, $h_i$ é o número total de chamadas originadas no nó de demanda $i$; $\bar{t}$ é tempo médio de duração de uma chamada atendida no sistema; $I$ é o número total de nós de demanda e $p$ é o número total de servidores do sistema.

É necessário também calcular, para cada área de demanda, o número $b$ de servidores que devem estar disponíveis a menos da distância crítica $S$ de forma a oferecer cobertura com confiabilidade α. Primeiro, devemos obter a expressão que considera a probabilidade de que um ou mais servidores estejam disponíveis para atender uma chamada no nó de demanda $i$ a menos de $S$ seja maior ou igual $\alpha$. Assim:

Pr[um ou mais servidores disponíveis a menos do padrão de resposta S] $\geq \alpha$
= (1- Pr[não há servidor disponível a menos de $S$ do nó $i$]) $\geq \alpha$
$$= 1 - \rho^{\sum_{k \in J} a_{ik} y_k} \geq \alpha \qquad \text{(A.9)}$$

Na expressão acima, $a_{ik}$ = 1 se área de demanda $i$ pode ser coberta por um servidor localizado em no nó $k$ ($a_{ik}$ = 0, caso contrário), e $y_k$ = 1 se um servidor é localizado no nó *k*. Desta forma, $\sum_{k \in J} a_{ik} y_k$ corresponde ao número de servidores disponíveis a menos de $S$ do nó de demanda $i$. Utilizando logaritmos, a expressão pode se tornar linear da forma:

$$\sum a_{ik} y_k \geq \frac{\log(1-\alpha)}{\log \rho}, \qquad \text{(A.10)}$$

O número de servidor $b$ é calculado através da seguinte expressão:

$\sum a_{ik} y_k \geq b$, onde $b = \left[ \frac{\log(1-\alpha)}{\log \rho} \right]$ sendo que [n] = menor inteiro maior ou igual a n.

Desta forma, para maximizar o número de chamadas com confiabilidade $\alpha$ é necessário maximizar o número de chamadas com ao menos $b$ servidores disponíveis a menos de $S$. Seja $x_{ij}$ = 1 se nó de demanda $i$ tem pelo menos $j$ servidores a menos de $S$ ($x_{ij}$ = 0, caso contrário), então $\sum_{j=1}^{b} x_{ij}$ corresponde ao número de vezes que o nó de demanda $i$ é coberta.

Assumindo que o número de vezes que o nó $i$ é coberto deve ser menor ou igual ao número de servidores disponíveis a menos de $S$ do nó $i$,podemos utilizar seguinte expressão para determinar se o nó $i$ é coberto com confiabilidade $\alpha$ :

$$\sum_{j=1}^{b} x_{ij} \leq \sum_{k \in J} a_{ik} y_k \tag{A.11}$$

O MALP I pode ser então ser formulado da seguinte forma:

$$\text{Maximize } Z = \sum_{i \in I} h_i x_{ib}, \tag{A.12}$$

$$\text{s.a } \sum_{j=1}^{b} x_{ij} \leq \sum_{k \in J} a_{ik}.y_k, i \in I \tag{A.13}$$

$$x_{ij} \leq x_{i(j-1),} i \in I, j = 2,..,b, \tag{A.14}$$

$$\sum_{k \in J} y_k = p \tag{A.15}$$

$$y_k, x_{ij} \in \{0,1\}, i \in I, j \in J, j = 2,..b. \tag{A.16}$$

A função objetivo busca maximizar as chamadas atendidas com confiabilidade $\alpha$. As restrições (A.13) asseguram que um nó de demanda $i$ tem cobertura com confiabilidade $\alpha$ se existirem pelo menos $b$ servidores a menos de $S$ da mesma. As restrições (A.14) estabelecem que para uma área de demanda seja coberta por $j$ servidores a menos de $S$, a mesma deve coberta por $j-1$ servidores, para $2 \leq j \leq b.$ A restrição (A.15) asseguram que $p$ servidores devem ser localizados e as restrições (A.16) estabelecem o caráter binário das variáveis de decisão.

### A3. Problema de Localização de Máxima Cobertura Esperada Ajustado (AMECLP - *Adjusted Maximal Expected Covering Location Problem*)

O modelo AMEXCL tem a formulação muito similar ao MEXCLP sendo que a função objetivo é modificada para:

$$\text{Maximizar } \sum_{i=1}^{I} \sum_{j=1}^{p} Q(p, \rho, j-1)(1-\rho)\rho^{j-1} h_i y_{ij} \tag{A.17}$$

onde $\rho$ = probabilidade de ocupação do servidor;

$I$ = número total de nós de demanda;

$p$ = número total de servidores localizados;

$h_i$ = demanda gerada no nó $i$;

$y_{ij}$ = 1 se o nó $i$ é coberto por ao menos por $j$ servidores; $y_{ij}$ = 0 caso contrário, e

$Q(p,\rho,j)$ = fatores de correção que relaxam a hipótese de independência entre os servidores. Estes fatores foram introduzidos por LARSON (1975) no modelo Hipercubo aproximado da seguinte forma:

$$Q(p,\rho,j)=\frac{\sum_{k=j}^{p-1}\left\{\left(\frac{(p-j-1)!(p-k)}{(k-j)!}\right)\left(\frac{p^p}{p!}\right)\rho^{k-j}\right\}}{\left[\left((1-\rho)\sum_{j=0}^{p-1}\left(\frac{p^j}{j!}\right)\rho^j\right)+\left(\frac{p^p\rho^p}{p!}\right)\right]} \qquad \text{(A.18)}$$

**A4. Problema de Localização de Máxima Disponibilidade Extendido (EMALP – *Extended Maximal Availibility Location Problem*)**

Maximizar $Z=\sum_{i\in I}h_i z_i$ (A.19)

s.a $\left[\left\{1-\prod_{j=1}^{p}\rho_j^{a_{ik}y_{jk}}Q\left(p,\rho,\sum_{j=1}^{p}\sum_{i\in I}a_{ik}y_{jk}-1\right)\right\}-\alpha\right]z_i\geq 0 \qquad \forall j\in J$ (A.20)

$$\sum_{k\in J}\sum_{j=1}^{p}y_{jk}=p\,, \qquad \text{(A.21)}$$

$$y_{jk},z_i\in\{0,1\},\forall k\in J,i\in I,j=1,2....,p. \qquad \text{(A.22)}$$

onde, além das definições anteriores, temos:

$\rho_j$ = fração de ocupação do servidor $j$;

$y_{jk}$ =1 se o servidor $j$ está localizado em $k$ e $y_{jk}$ = 0 caso contrário;

$z_i$ =1 se a área de demanda $i$ tem pelo menos um servidor disponível a menos de S com confiabilidade $\alpha$ e $z_i$ =0 caso contrário.

A função objetivo busca maximizar a cobertura total obtida com a localização dos $p$ servidores. As restrições (A.20) asseguram que cada nó de demanda $i$ está coberto com confiabilidade $\alpha$, dado que o número de servidores do sistema é $p$. A restrição (A.21) assegura que número de servidores localizados é $p$ e as restrições (A.22) estabelecem o caráter binário das variáveis de decisão.