

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

REGIANE MÁXIMO DE SOUZA

**ANÁLISE DA CONFIGURAÇÃO DE SAMU UTILIZANDO MODELO
HIPERCUBO COM PRIORIDADE NA FILA E MÚLTIPLAS
ALTERNATIVAS DE LOCALIZAÇÃO DE AMBULÂNCIAS**

SÃO CARLOS 2010

**ANÁLISE DA CONFIGURAÇÃO DE SAMU UTILIZANDO MODELO
HIPERCUBO COM PRIORIDADE NA FILA E MÚTIPLAS
ALTERNATIVAS DE LOCALIZAÇÃO DE AMBULÂNCIAS**

**UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO**

**ANÁLISE DA CONFIGURAÇÃO DE SAMU UTILIZANDO MODELO
HIPERCUBO COM PRIORIDADE NA FILA E MÚLTIPLAS
ALTERNATIVAS DE LOCALIZAÇÃO DE AMBULÂNCIAS**

Regiane Máximo de Souza

**Tese apresentada ao Programa de Pós-
Graduação em Engenharia de Produção
da Universidade Federal de São Carlos,
como parte dos requisitos para obtenção
do Título de Doutora em Engenharia de
Produção.**

**ORIENTADOR: Prof. Dr. Reinaldo Morabito.
COORIENTADOR: Prof. Dr. Fernando Y. Chiyoshi.**

**SÃO CARLOS
2010**

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

S729ac

Souza, Regiane Máximo de.

Análise da configuração de SAMU utilizando modelo hipercubo com prioridade na fila e múltiplas alternativas de localização de ambulâncias / Regiane Máximo de Souza. -- São Carlos : UFSCar, 2010.
221 f.

Tese (Doutorado) -- Universidade Federal de São Carlos, 2010.

1. Pesquisa operacional. 2. Processos estocásticos (Aplicações). 3. Teoria das filas. 4. Modelo hipercubo. 5. Atendimento médico - emergencial. I. Título.

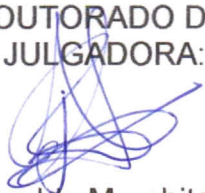
CDD: 658.4034 (20^a)



FOLHA DE APROVAÇÃO

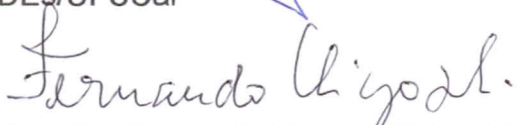
Aluno(a): Regiane Máximo de Souza

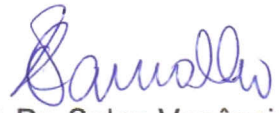
TESE DE DOUTORADO DEFENDIDA E APROVADA EM 23/08/2010 PELA
COMISSÃO JULGADORA:

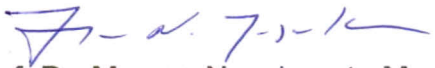

Prof. Dr. Reinaldo Morabito Neto
Orientador(a) PPGE/UFSCar

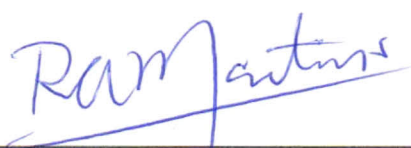

Prof. Dr. Roberto Antonio Martins
PPGE/UFSCar


Prof. Dr. Francisco Louzada Neto
DEs/UFSCar


Prof. Dr. Fernando Yassuo Chiyoshi
COPPE/UFRJ


Prof. Dr. Solon Venâncio de Carvalho
INPE/LAC


Prof. Dr. Marcos Nascimento Magalhães
IME/USP


Prof. Dr. Roberto Antonio Martins
Coordenador do PPGE

Dedico este trabalho aos meus pais, pelo apoio
incondicional neste e em todos os
projetos de minha vida.
Aos meus filhos, sem eles não sei se teria
seguido em frente.

“O homem é do tamanho do seu sonho”
Fernando Pessoa

AGRADECIMENTOS

Ao meu orientador Reinaldo Morabito, que me orientou com seriedade, disciplina e muita competência. Conduziu-me sabiamente em vários pontos críticos do trabalho. Além disso, agradeço pela confiança, incentivo, paciência, dedicação e amizade em todos estes anos da realização desta tese.

Ao meu co-orientador Fernando Y. Chiyoshi, por ter acreditado neste projeto. Pelas horas, durante meses, que ficamos reunidos pelo Skype[®] e por sua paciência e amizade. Sua contribuição, com certeza, foi além do escopo desta tese.

À Ana Paula Ianonni, sua contribuição, amizade e dedicação foram fundamentais para a realização da simulação no ARENA[®] contida nesta pesquisa.

Ao SAMU-RP pelo apoio e valiosa colaboração durante o desenvolvimento deste trabalho, auxiliando a coleta de dados, na qual se tornou possível a realização do estudo de caso desta tese.

Aos membros da banca de qualificação (prof. Solon Venâncio de Carvalho, prof. Marcos Nascimento Magalhães e prof. José Gilberto Rinaldi) pelas sugestões e contribuições ao desenvolvimento desta pesquisa.

Às pessoas importantes da minha vida: meus pais, meus filhos e o Rui, que sempre me incentivou e acompanhou todas as minhas fases de produção desta tese.

Aos meus amigos de Franca, inclusive os “importados!!”, em especial ao Maurício Buffa e Neiva, que durante estes anos sempre me incentivaram em nossos famosos “furdunços”.

SUMÁRIO

Lista de Figuras.....	i
Lista de Tabelas	iii
Lista dos principais símbolos e siglas utilizados	vii
Resumo	ix
<i>Abstract</i>	xi
1. Introdução	1
1.1 Objetivos do trabalho	7
1.2 Organização do texto	8
2. <i>Sistemas de Atendimento Emergencial e Modelos de Localização que Utilizam Teoria das Filas</i>.....	11
2.1 Sistemas de Atendimento Emergencial	11
2.1.1 Sistemas de Atendimento Móvel de Urgência (SAMU's)	15
2.1.2 O SAMU de Ribeirão Preto (SAMU-RP).....	18
2.2 Modelos de localização.....	22
2.2.1 Modelos que utilizam Teoria das filas	23
2.2.2 Modelos que combinam o modelo hipercubo com procedimentos de otimização.....	26
2.2.3 Modelos de localização dinâmica	27
2.2.4 Modelos de simulação.....	28
3. <i>Teoria das Filas com Distribuições Exponenciais e Modelo Hipercubo</i>	31
3.1 O processo de Poisson e a distribuição exponencial	31
3.2 A fórmula de Little	35
3.3 A fila $M/M/m$	36
3.4 A fila $M/M/m/C$	41
3.5 O modelo hipercubo clássico.....	42
3.6 Filas com classes de usuários e distribuições exponenciais.....	52
3.6.1 Prioridades sem interrupção em um sistema $M_p/M/m$	53
3.6.2 Prioridades sem interrupção no modelo hipercubo	56
4 <i>Extensão do modelo hipercubo para análise do SAMU-RP</i>	59

4.1	Exemplo ilustrativo	59
4.1.1	Transição de estados do modelo ilustrativo:.....	63
4.1.2	Transição de estados da fila S :.....	70
4.2	Medidas de Desempenho do Exemplo Ilustrativo	73
4.3	Resultados computacionais e relação do modelo hipercubo com prioridade com o modelo $M/M/m/C$	83
4.3.1	Medidas de desempenho do modelo hipercubo com prioridade para servidores homogêneos	85
4.3.2	Medidas de desempenho do modelo $M/M/m/C$ e comparação com o modelo hipercubo com prioridade do exemplo ilustrativo	89
4.3.3	Medidas de desempenho do modelo $M_r/M/m$ e do modelo hipercubo com prioridade do exemplo ilustrativo	91
4.3.4	Modelo hipercubo com prioridade com servidores heterogêneos e comparação com o modelo de simulação do exemplo ilustrativo.....	92
4.4	Generalização das equações da fila do modelo hipercubo considerando prioridade na fila.....	98
5	<i>Aplicação do modelo hipercubo no SAMU-RP e análise dos resultados</i>	105
5.1	Descrição dos atendimentos	105
5.2	Períodos de pico.....	108
5.3	Validação das hipóteses do modelo hipercubo para o SAMU-RP	109
5.3.1	Área dividida em N_A átomos geográficos	110
5.3.2	Processo de chegada	110
5.3.2.1	Cálculo dos λ_{jk}	113
5.3.3	Tempos de viagem.....	115
5.3.4	Servidores.....	116
5.3.5	Localização dos servidores.....	116
5.3.6	Despacho dos servidores	117
5.3.7	Política de despacho dos servidores.....	118
5.3.8	Tempos de atendimento.....	120
5.3.9	Relação entre o tempo de atendimento e o tempo de viagem.....	122
5.4	Resultados da aplicação do modelo hipercubo	123
5.4.1	Resultados para o SAMU-RP (cenário original)	124
5.4.2	Cenário 1 – Atendimento dos chamados de remoção	132
5.4.3	Cenário 2 – Aumento de demanda	137
5.4.4	Cenário 3 – Período da manhã com uma ambulância a menos	142

5.4.5	Cenário 4 – Período da tarde com uma ambulância a menos.....	144
5.4.5	Cenário 5 – Período da noite com uma ambulância a menos.....	146
6	Conclusões	149
6.1	Perspectivas de Pesquisa Futura	152
	Referências Bibliográficas.....	155
	Anexos	165

LISTA DE FIGURAS

Figura 2.1 – Distribuição espacial do sistema SAMU-Campinas no final dos anos 90.	17
Figura 2.2 – Ambulâncias VSB's do SAMU-RP.	19
Figura 2.3 – Equipamentos dos veículos VSB's do SAMU-RP.	20
Figura 2.4 – Equipamentos do VSA no SAMU-RP.	21
Figura 2.5 – Distribuição espacial do sistema SAMU - Ribeirão Preto em 2005.	21
Figura 3.1 – Atendimento de três usuários no sistema.	35
Figura 3.2-Transições de estado do modelo $M/M/m$ a partir do estado n , quando $n < m$.	37
Figura 3.3 - Transições de estado do modelo $M/M/m$ a partir do estado n , quando $n \geq m$.	38
Figura 3.4- Transição de estados do modelo $M/M/m$.	40
Figura 3.5 - Transição de estados do modelo $M/M/m/c$.	41
Figura 3.6 – Estados do sistema com três servidores.	43
Figura 3.7 – Representação de um sistema de filas com r classes e disciplina de prioridade.	53
Figura 4.1 – Átomos geográficos	59
Figura 4.2– Estados do sistema com três servidores.	62
Figura 4.3 – Vértice $\{101\}$ e seus adjacentes.	64
Figura 4.4 – Vértice $\{011\}$ e seus adjacentes.	65
Figura 4.5 – Espaço de estados da fila 3 classes de usuários e com prioridades.	68
Figura 4.6– Vértice $\{ab\}$ e seus adjacentes.	70
Figura 4.7 – Vértice $\{ac\}$ e seus adjacentes.	71
Figura 4.8 – Exemplo de geração dos estados da fila para $r = 3$.	100
Figura 4.9 – Matriz gerada pelos procedimentos de generalização das equações de balanço da fila.	104
Figura 5.1 – atendimentos de urgência e emergência realizados pelo SAMU_RP em 2005.	106
Figura 5.2 – atendimentos de urgência e emergência realizados pelo SAMU_RP em agosto de 2005.	107
Figura 5.3 – atendimentos realizados pelo SAMU_RP no 4º dia observado em agosto de 2005.	107

Figura 5.4 – Átomos geográficos do SAMU_RP.....	110
Figura 5.5 - Gráfico obtido no BestFit do teste de aderência do processo de chegada, para o período da manhã.....	112
Figura 5.6 - Gráfico obtido no BestFit do teste de aderência do processo de chegada, para o período da tarde.	112
Figura 5.7 - Gráfico obtido no BestFit do teste de aderência do processo de chegada, para o período da noite.	112
Figura 5.8 - Porcentagem do número de chamados em cada subátomo nos períodos manhã, tarde e noite.....	113
Figura 5.9 – Localização das ambulâncias no SAMU-RP.	124
Figura 5.10 – <i>Workloads</i> das ambulâncias.	125
Figura 5.11 – <i>Workloads</i> das ambulâncias para o cenário 1.	133
Figura 5.12 – Localização das ambulâncias no cenário 3.	142
Figura 5.13 – Localização das ambulâncias no cenário 4.	144
Figura 5.14 – Localização das ambulâncias no cenário 5.	146

LISTA DE TABELAS

Tabela 2.1 – Quantidade de acidentes por tipo de dias da semana.....	18
Tabela 3.1– Matriz de Preferências de despacho.	46
Tabela 4.1 – Lista de preferências de despacho	61
Tabela 4.2 – Matriz do tempo de viagem do servidor i ao subátomo j , t_{ij}	62
Tabela 4.3 – Taxas de chegada (λ) dos átomos (última linha), das classes (última coluna) e dos subátomos (células internas).	84
Tabela 4.4 – Matriz dos tempos de viagem entre os átomos, $\tau_{il,jk}$	85
Tabela 4.5 – Matriz da localização dos servidores, $l_{m,jk}$	85
Tabela 4.6 – Probabilidade de estado do exemplo ilustrativo.	86
Tabela 4.7 – <i>Workloads</i> das ambulâncias.....	86
Tabela 4.8 – Frequências de despacho do servidor i para o subátomo jk	86
Tabela 4.9 – Frequências de despacho <i>backup</i> do servidor i	87
Tabela 4.10 – Frequências de despacho <i>backup</i> do subátomo jk	87
Tabela 4.11 – Tempo médio de espera na fila, W_q , W_{qa} , W_{qb} e W_{qc}	87
Tabela 4.12 – Tempo médio de viagem e resposta no sistema.....	88
Tabela 4.13 – Tempo médio de espera na fila para os tipos de subátomos a , b e c	88
Tabela 4.14 – Tempo médio de viagem e de resposta ao subátomo jk	88
Tabela 4.15 – Tempo médio de viagem e de resposta aos tipos de subátomos a , b e c	89
Tabela 4.16 – Tempo médio de viagem e de resposta do servidor i	89
Tabela 4.17 – Tempo de viagem e de resposta do servidor i para os tipos de subátomos a , b e c	89
Tabela 4.18 – Relação das probabilidades dos estados do modelo $M/M/m/C$ com os estados do modelo hipercubo.).....	90
Tabela 4.19 – Comparação das medidas de desempenho do modelo $M/M/m/C$ com o modelo hipercubo.	91
Tabela 4.20 – Comparação do modelo hipercubo com o modelo $M_r/M/m$	91
Tabela 4.21 – Taxa de serviço para os três servidores.	92
Tabela 4.22 – Probabilidade de estado.	92
Tabela 4.23 – <i>Workloads</i> das ambulâncias.....	93
Tabela 4.24 – Frequências de despacho do servidor i para o subátomo j	93
Tabela 4.25 – Frequências de despacho <i>backup</i> do servidor i	93

Tabela 4.26 – Frequências de despacho <i>backup</i> do subátomo <i>jk</i> .	94
Tabela 4.27 – Tempo médio de espera na fila W_q , W_{qa} , W_{qb} e W_{qc} .	94
Tabela 4.28 – Tempo médio de viagem e resposta no sistema.	94
Tabela 4.29 – Tempo médio de viagem e de resposta ao subátomo <i>jk</i> .	95
Tabela 4.30 – Tempo médio de viagem e de resposta para as classes <i>a</i> , <i>b</i> e <i>c</i> .	95
Tabela 4.31 – Tempo médio de viagem e de resposta do servidor <i>i</i> .	95
Tabela 4.32 – Tempo de viagem e de resposta do servidor <i>i</i> para as classes <i>a</i> , <i>b</i> e <i>c</i> .	96
Tabela 4.33 – <i>Workload</i> dos servidores para até 5 usuários na fila.	96
Tabela 4.34 – Tempo médio de espera na fila W_q , W_{qa} , W_{qb} e W_{qc} , para até 5 usuários na fila.	96
Tabela 4.35 – Tempo de espera na fila do modelo hipercubo clássico, considerando prioridade na fila.	97
Tabela 4.36 – Tempo de viagem e de resposta do modelo hipercubo clássico, considerando prioridade na fila.	98
Tabela 5.1 – Análise do período de pico para os três períodos do dia: manhã, tarde e noite.	109
Tabela 5.2 – Proporção de chamados em cada subátomo do sistema.	111
Tabela 5.3 – Intervalos médios entre chegadas sucessivas no sistema, no período da noite.	114
Tabela 5.4 – Taxas médias de chegada dos chamados (por hora) para cada subátomo.	115
Tabela 5.5 – Tempo médio de viagem entre subátomos obtidos a partir de dados do SAMU-RP.	116
Tabela 5.6 – Matriz de localização dos servidores nos subátomos para os períodos da manhã e tarde.	117
Tabela 5.7 – Matriz de localização dos servidores nos subátomos para o período da noite.	117
Tabela 5.8 – Exemplo: matriz de preferência de despachos no cenário original.	120
Tabela 5.9 – Tempos e taxas médias de atendimento para cada ambulância.	145
Tabela 5.10 - Relação entre o tempo de atendimento e o tempo de viagem para as ambulâncias.	122
Tabela 5.11 – Tempo médio de espera na fila.	126
Tabela 5.12 – <i>Workload</i> das ambulâncias.	126
Tabela 5.13 – Tempo médio de viagem e de resposta no sistema.	127

Tabela 5.14 – Tempos médios de viagem das ambulâncias.....	127
Tabela 5.15 – Tempos médios de viagem para cada subátomo.....	128
Tabela 5.16 – Tempo médio de viagem nas classes a , b e c	128
Tabela 5.17 - Tempo médio de viagem e de resposta de cada ambulância nos átomos a , b e c	129
Tabela 5.18 – Tempos médios de resposta das ambulâncias.....	129
Tabela 5.19– Tempos médios de resposta para cada subátomo.	130
Tabela 5.20 – Tempo médio de resposta nas classes a , b e c	130
Tabela 5.21 – Taxas médias de chegada dos chamados para cada átomo para o cenário 1.	133
Tabela 5.22 – Tempo médio de espera em fila para o cenário 1.	134
Tabela 5.23 –Tempo médio de viagem e de resposta no sistema para o cenário 1.	134
Tabela 5.24 – <i>Workload</i> das ambulâncias do cenário 1.	134
Tabela 5.25 – Tempos médios de viagem e de resposta das ambulâncias para o cenário 1.	135
Tabela 5.26 – Tempos médios de viagem e de resposta para cada átomo para o cenário 1.	135
Tabela 5.27 – Tempo médio de viagem e resposta para os átomos a , b e c para o cenário 1.	136
Tabela 5.28 - Tempo médio de viagem de cada ambulância para as classes a , b e c para o cenário 1.	136
Tabela 5.29 – <i>Workloads</i> do cenário 2 - Aumento de demanda no período da manhã.	138
Tabela 5.30 – Tempos médios de espera na fila do cenário 2 - Aumento de demanda no período da manhã.	138
Tabela 5.31 – Tempos médios de resposta para os subátomos do cenário 2 - Aumento de demanda no período da manhã.	139
Tabela 5.32 – <i>Workloads</i> do cenário 2 - Aumento de demanda no período da tarde... ..	140
Tabela 5.33 – Tempos médios de espera na fila do cenário 2 - Aumento de demanda no período da tarde.	140
Tabela 5.34 – Tempos médios de resposta para os subátomos do cenário 2 - Aumento de demanda no período da tarde.....	141
Tabela 5.35 - <i>Workloads</i> do cenário 3.....	142
Tabela 5.36 – Tempos médios de espera na fila do cenário 3.	143
Tabela 5.37 – Tempo médio de resposta para os subátomos do cenário 3.....	143

Tabela 5.38 – <i>Workloads</i> do cenário 4.	145
Tabela 5.39 – Tempos médios de espera na fila do cenário 4.	145
Tabela 5.40 – Tempo médio de resposta do cenário 4.	146
Tabela 5.41 – <i>Workloads</i> do cenário 5.	147
Tabela 5.42 – Tempos médios de espera na fila do cenário 5.	147
Tabela 5.43 – Tempo médio de resposta do cenário 5.	148

LISTA DOS PRINCIPAIS SÍMBOLOS E SIGLAS UTILIZADOS

b_i	estado do servidor i ;
B	representação vetorial de um estado do sistema (i.e., $B = \{b_1, b_2, \dots, b_N\}$);
E_{ij}	conjunto dos estados nos quais o servidor i é o primeiro servidor disponível na lista de despacho do átomo j ;
E_{ijk}	conjunto dos estados nos quais o servidor i é o primeiro servidor disponível na lista de despacho do subátomo jk ;
EMS	<i>Emergency Medical Systems</i> ;
$f_{ij}^{[nq]}$	fração de despachos do servidor i que é enviado ao átomo j que implicam tempo de espera em fila;
$f_{ij}^{[q]}$	fração de despachos do servidor i que é enviado ao átomo j que não implicam tempo de espera em fila;
$f_{ijk}^{[nq]}$	fração de despachos do servidor i que é enviado ao subátomo jk que implicam tempo de espera em fila;
$f_{ijk}^{[q]}$	fração de despachos do servidor i que é enviado ao subátomo jk que não implicam tempo de espera em fila;
f_{ij}	fração de despachos no sistema que são atendidas pelo servidor i no átomo j ;
L_{ij}	matriz de localização do servidor i no átomo j ;
L_{ijk}	matriz de localização do servidor i no subátomo jk ;
L_q	número médio de usuários na fila;
L_{qk}	número médio de usuários na fila da classe k ;
λ	taxa total de chegada no sistema;
λ_j	taxa de chegada de chamadas no átomo j ;
λ_p	taxa de chegada dos chamados de qualquer átomo, com prioridade p ;
μ	taxa total de atendimento no sistema;
μ_i	taxa de atendimento do servidor i ;
N	número de servidores no sistema;
n	índice em geral utilizado para denotar o número de usuários no sistema;
N_A	número de átomos no sistema;
p	classes de usuários no sistema;

P	número de classes de usuários do sistema;
P_B	probabilidade de equilíbrio estado B ;
P_Q	probabilidade de saturação do sistema menos a probabilidade de todos os servidores estarem ocupados;
P_{Qk}	soma dos estados da fila onde há pelo menos um chamado da classe k ;
P_S	probabilidade de saturação do sistema;
ρ	carga média de trabalho no sistema;
ρ_i	carga de trabalho do servidor i ;
S_p^{n-N}	estado onde há $n - N$ usuários em fila com prioridade p ;
SAE	sistemas de Atendimento Emergencial;
SAMU	sistema de Atendimento Móvel de Urgência;
\bar{T}	tempo médio de viagem no sistema;
\overline{TU}_i	tempo médio de viagem do servidor i ;
\overline{TU}_{ik}	tempo médio de viagem do servidor i atendendo usuários da classe k ;
\bar{T}_j	tempo médio de viagem ao átomo j ;
\bar{T}_{jk}	tempo médio de viagem ao subátomo jk ;
t_{ij}	matriz dos tempos médios de viagem do servidor i ao átomo j ;
t_{ijk}	matriz dos tempos médios de viagem do servidor i ao subátomo jk ;
\bar{T}_Q	tempo médio de viagem para chamados em fila;
\overline{TU}_n	tempo médio de viagem do servidor n ;
τ_{jp}	matriz dos tempos médios de viagem entre os átomos p e j ;
$\tau_{jk,pl}$	matriz dos tempos médios de viagem entre os subátomos pl e jk ;
UTI	veículo com equipamentos médicos emergenciais especializados;
VSA	Veículo de Suporte Avançado;
VSB	Veículo de Suporte Básico;
W_q	tempo médio de espera na fila;
W_{qk}	tempo médio de espera na fila para usuários da classe k .

RESUMO

Em alguns Sistemas de Atendimento Emergenciais a demanda pelo serviço é alta devido ao atendimento a pacientes em estado grave a leve, fazendo aumentar o nível de utilização dos servidores. Nesses sistemas, pode haver formação de filas de espera e a necessidade de considerar explicitamente políticas de prioridade no atendimento é extremamente importante e requer extensões no modelo hipercubo que nunca foram exploradas na literatura. Ainda, a demanda geográfica e temporal dos SAMU's pode mudar ao longo do dia devido a sua natureza aleatória. Os objetivos do presente estudo são: (a) estender o modelo hipercubo para considerar fila com prioridade, o que até onde se tem conhecimento nunca foi feito na literatura e (b) propor uma abordagem para múltiplas configurações de localização das ambulâncias, explorando variações importantes da demanda e do serviço ao longo do dia. Para verificar a viabilidade e a aplicabilidade desta abordagem, é realizado um estudo de caso no SAMU de Ribeirão Preto-SP (SAMU-RP) que, além dos atendimentos de urgência e emergência, opera atendendo remoção de pacientes (transporte de pacientes entre hospitais, de domicílio para hospital ou vice-versa). Além da configuração original do SAMU-RP, foram analisados cinco cenários alternativos que consideram questões importantes: o impacto dos atendimentos de remoção; o impacto do aumento da demanda no período mais congestionado e a possibilidade de múltiplas configurações de localização das ambulâncias nos três períodos analisados em importantes medidas de desempenho do sistema, tais como *workloads* das ambulâncias, tempos de viagem das ambulâncias, tempos de resposta aos usuários, entre outros. Os resultados mostram que o modelo hipercubo, estendido para tratar prioridade na fila, pode ser utilizado para analisar satisfatoriamente sistemas como o SAMU-RP, permitindo uma avaliação suficientemente rápida e precisa do desempenho do sistema em diversos cenários.

ABSTRACT

In some emergency medical systems the service demand is high due to the treatment of patients in the range severe to mild, which increases the utilization level of the servers. In these systems, may be queues formation and so the need to explicitly consider priority in care is extremely important. In this study we extend the hypercube model to deal with this situation never explored in the literature. Besides that, the geographical and temporal demand of SAMU can change throughout the day due to their random nature. The goals of this study are: (a) to extend the model in order to consider hypercube priority queue, which as far as we know has never been done in the literature and (b) to propose an approach for multiple configurations of ambulances' localization, exploring important variations in demand and service throughout the day. In this work, in order to verify the feasibility and applicability of this approach, we conducted a case study at Ribeirão Preto's SAMU (SAMU-RP) that, apart from urgent care and emergency operates removing patients (transporting patients between hospitals, from hospital to home and vice versa). Besides the original configuration of SAMU-RP, we analyzed five alternatives scenarios to examine three important issues: the impact of the removals, the impact of increased demand at the busiest periods and the possibility of multiple configurations for the localization of ambulances in the three analyzed periods in important performance measures of the system, like: workloads, travel times, response times for users, etc. The results show that the hypercube model with priority in the queue can be used to analyze systems like SAMU-RP and it allows sufficiently rapid and accurate evaluation of the performance, of the system in several scenarios.

1. Introdução

A qualidade de vida da população está diretamente ligada ao acesso à saúde. Principalmente em áreas urbanas, há um grande número de acidentes e outras ocorrências de urgência e emergência, como infarto, intoxicação, queimadura, afogamento e queda accidental. No Brasil, há superlotações nos hospitais e prontos-socorros, levando a baixa qualidade no nível de resposta do sistema de saúde às urgências e emergências, o que provoca grandes esperas nesses locais de atendimento. Nesse contexto, a organização dos atendimentos emergenciais ganham maior relevância e causam forte impacto ao setor saúde. A resposta rápida a tal demanda é fundamental para minimizar possíveis sequelas decorrentes no quadro dos pacientes.

No Brasil, em 2007, mais de 35.000 pessoas perderam a vida vítimas de acidentes de trânsito (OPAS, 2008), porém acredita-se que esses números sejam ainda maiores. Só nas rodovias federais, ocorreram 128.456 acidentes, sendo 5.757 acidentes com mortes e 75.462 sem vítimas (DNIT, 2009). Em todo o mundo, o trânsito causa perda de vidas, mas os números brasileiros são alarmantes. Ainda em 2007, conforme a OMS (2009), o país ocupou o 5º lugar em mortes no trânsito no mundo.

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

Em 29 de setembro de 2003, entraram em vigor duas importantes portarias, a Portaria nº 1863 GM/MS, que instituiu a Política Nacional de Atenção às Urgências, a qual tem como um de seus componentes o atendimento pré-hospitalar móvel, por meio do Parecer 15/98 do Conselho Federal de Medicina, e a Portaria nº 1864 GM/MS, que oficializou a implantação do Serviço de Atendimento Móvel de Urgência (SAMU-192)

em alguns municípios e regiões do território brasileiro. No Brasil, o SAMU (*Service d'Aide Médicale d'Urgence*, de origem francesa) teve início a partir de um acordo bilateral assinado entre Brasil e a França, por solicitação do Ministério da Saúde. As viaturas de suporte avançado possuem obrigatoriamente a presença do médico, diferentemente dos moldes norte-americanos em que as atividades de resgate são exercidas primariamente por profissionais paramédicos (esse profissional não existe no Brasil) (LOPES e FERNANDES, 1999).

O SAMU é um programa do governo federal que tem a finalidade de prestar socorro emergencial às pessoas e garantir a qualidade no atendimento. Assim, é um grande desafio para o poder público oferecer um serviço de boa qualidade para a população, com as restrições dos recursos disponíveis. O serviço funciona 24 horas por dia e é composto por profissionais da saúde: médicos, enfermeiros, auxiliares de enfermagem e socorristas. Esse serviço presta o atendimento em qualquer local, residências, locais de trabalho e vias públicas. O socorro é feito após uma ligação telefônica gratuita ao número 192. Nesse contexto, os SAMU's têm um papel importante na sociedade.

O SAMU foca cinco grandes ações:

- organizar o atendimento de urgência nos prontos-atendimentos, unidades básicas de saúde e nas equipes do Programa Saúde da Família;
- estruturar o atendimento pré-hospitalar móvel (SAMU 192);
- reorganizar as grandes urgências e os prontos-socorros em hospitais;
- criar a retaguarda hospitalar para os atendimentos nas urgências;
- reestruturar o atendimento pós-hospitalar.

Em 2008, a rede Nacional SAMU 192 possuía 147 serviços de atendimento móvel às urgências, atendendo 1.273 municípios brasileiros (IBGE, 2008), oferecendo o serviço a um total de 112 milhões de pessoas para uma população brasileira estimada de 189 milhões, quase dois terços da população brasileira (IBGE, 2008 e MS, 2008).

Os SAE's, como os SAMU's, são caracterizados essencialmente por incertezas quanto a disponibilidade, localização, tempo de serviço dos servidores, demanda ao longo da região e tempo de resposta para atendimento aos usuários. Os sistemas de atendimento a emergências em saúde caracterizam um grande desafio para

todas as nações, pois independentemente do tipo de urgência envolvida, somente com uma rigorosa organização é possível oferecer um serviço de boa qualidade. Em sistemas médico-emergenciais, o tempo de resposta ao usuário é uma das principais medidas de desempenho. Como enfatizado antes, o atraso no atendimento dos chamados nesses sistemas está relacionado ao conflito entre a demanda por serviço e as restrições de capacidade do sistema. Nos últimos 50 anos, vem aumentando o interesse de pesquisadores em estudos de sistemas emergenciais em serviços urbanos (SIMPSON e HANCOCK, 2009).

Algumas medidas de desempenho mais relevantes para um sistema de atendimento de urgência podem ser divididas em: medidas externas, do ponto de vista do usuário, como o tempo médio de resposta a um chamado, o tempo médio de viagem para cada área da cidade (referida neste estudo como átomo) e a frequência de chamadas atendidas em um tempo inferior a um limite determinado; e medidas internas, do ponto de vista do gerente do sistema, como a carga de trabalho das ambulâncias, as frequências de despacho das ambulâncias para os átomos, a fração de atendimentos realizados fora da área de cobertura de cada ambulância e o tempo médio de viagem para cada ambulância.

O chamado modelo hipercubo, proposto originalmente por Larson (1974) e baseado na teoria de filas espacialmente distribuídas, tem se mostrado eficiente e preciso para analisar SAE's, como foi analisado, por exemplo, nos Estados Unidos, em Chelst e Barlach (1981), Brandeau e Larson (1986), Burwell *et al.* (1993), Sacks e Grief (1994), Swersey (1994) e Larson e Odoni (2007). No Brasil, alguns exemplos aparecem em Gonçalves *et al.* (1994), Gonçalves *et al.* (1995), Mendonça (1999), Mendonça e Morabito (2000), Oliveira (2003), Chiyoshi *et al.* (2000), Costa *et al.* (2004), Figueiredo *et al.* (2005), Takeda *et al.* (2004, 2007) e Iannoni (2005). A aplicação original do modelo hipercubo foi desenvolvida para o problema de patrulhamento policial, mas depois o modelo passou a ser aplicado em vários sistemas de emergência, como empresas de segurança, bombeiros, ambulâncias, reparos em redes de energia elétrica etc. (LARSON e ODONI, 2007). Alguns exemplos no setor de saúde no Brasil podem ser vistos em Gonçalves *et al.* (1994), Gonçalves *et al.* (1995), Takeda (2000), Takeda *et al.* (2004, 2007), Mendonça e Morabito (2000, 2001), Chiyoshi *et al.* (2000), Figueiredo *et al.* (2005) e Luque (2006).

O modelo hipercubo tem por objetivo avaliar a configuração e estimar as medidas de desempenho para SAE's, como, por exemplo, os SAMU's, possibilitando um planejamento adequado e melhores níveis de serviço oferecido. Em sua forma original, trata-se de um modelo descritivo utilizado para analisar e planejar um sistema emergencial, com enfoque não apenas em análise e cálculo das medidas de desempenho do sistema, mas também de otimizá-lo. Portanto, é utilizado uma ou mais medidas de desempenho, é necessário combinar o modelo hipercubo, que é essencialmente descritivo, com procedimentos de otimização, como, por exemplo, heurísticas construtivas, algoritmos genéticos e busca tabu (GALVÃO e MORABITO, 2008).

A partir da combinação do modelo hipercubo com procedimentos de otimização é possível tratar o dimensionamento das áreas de cada servidor, de forma que otimize a configuração e a operação do sistema, em termos das medidas de desempenho mais relevantes, tanto do ponto de vista do usuário, como dos operadores do sistema. Dessa forma, é viável estudar o *trade-off* de medidas conflitantes, como, por exemplo, tempo médio de resposta ao usuário e a carga de trabalho das ambulâncias. São poucas e, relativamente recentes, as publicações com o enfoque de integração do modelo hipercubo com procedimentos de otimização. Algumas aplicações podem ser vistas em Batta *et al.* (1989), Saydam e Aytug (2003), Chiyoshi *et al.* (2003), Galvão *et al.* (2003, 2005, 2008) e Iannoni *et al.* (2006, 2008a, 2008b).

Muitos SAMU's, como o da cidade de Campinas-SP, trabalham com duas ou mais classes de usuários, separando os chamados em emergências e urgências. Para esses sistemas é necessário incorporar políticas de prioridade no modelo. Os SAMU's possuem particularidades que os diferem entre si e que devem ser estudadas e consideradas no tratamento do modelo, como diferenças com relação a tipo, número e localização das ambulâncias, regras de despacho, tipos de chamados, duplo despacho, *bachup* parcial etc. Takeda (2000) e Takeda *et al.* (2004, 2007) estudaram o SAMU-Campinas. Na época do estudo, o sistema operava com todas as ambulâncias centralizadas e atendia duas classes de usuários: as básicas, atendidas prioritariamente pelas VSB's (Veículos de Suporte Básico) e as avançadas, atendidas prioritariamente pelas VSA's (Veículos de Suporte Avançado). O sistema permitia formação de fila com usuários de até, no máximo, o número de servidores disponíveis em operação do sistema. O estudo mostrou que, mesmo sem considerar prioridade na fila, o modelo hipercubo foi adequado para analisar sistemas como o SAMU-Campinas e, ainda,

verificou-se a possibilidade de descentralização das ambulâncias por meio da avaliação de diversos cenários alternativos.

Diferentemente do SAMU-Campinas, o SAMU da cidade de São Paulo opera com um grande número de ambulâncias, o que causa uma dificuldade para sua análise por meio do modelo hipercubo, em função dos requisitos computacionais para a resolução do modelo, como visto adiante. O número de estados do modelo aumenta exponencialmente com o aumento do número de ambulâncias [$O(2^m)$, sendo m o número de ambulâncias]. Como consequência, há um aumento exponencial no número de variáveis e equações de balanço do modelo hipercubo, dificultando significativamente sua resolução. Esse problema foi estudado por Luque (2006).

Em Ribeirão Preto, o SAMU foi criado a partir de iniciativa de profissionais da Secretaria da Saúde, tendo como princípios o atendimento às urgências no campo pré-hospitalar, que são organizadas pela sua central de regulação. O SAMU de Ribeirão Preto entrou em operação no dia 08 de outubro de 1996, após um longo período de idealização e adequação. O serviço foi constituído por uma equipe de suporte avançado, já se prevendo, para o futuro, mudanças importantes no projeto inicialmente viabilizado. Essa equipe, constituída pelos elementos obrigatórios, - médico, enfermeira e motorista - conquistou o seu lugar no atendimento emergencial pré-hospitalar no município, atividade até então exclusiva da equipe de resgate do Corpo de Bombeiros. Nesse período, foi dimensionada a real função do SAMU frente à população local e às autoridades competentes, vinculando, de forma definitiva, o atendimento médico-emergencial ao paciente crítico, neste momento em ambiente pré-hospitalar.

A cidade de Ribeirão Preto-SP possui uma estimativa populacional de 558 mil habitantes para 2008 (estimativa IBGE, 2008), com tendência de crescimento populacional na ordem de 2,5% ao ano, conforme estimativas do próprio IBGE. Com o aumento populacional, o número de carros tem aumentado significativamente na cidade e a ocorrência de acidentes de trânsito também vem aumentando nos últimos anos.

Para os propósitos de modelagem de um SAMU, um aspecto importante é o atendimento a diferentes classes de usuários. No SAMU-RP, similarmente a alguns outros SAMU's, são atendidas três classes de usuários: chamados graves (emergências), atendidos prioritariamente pelo VSA (Veículo de Suporte Avançado); chamados

moderados (urgências 1) e chamados leves (urgências 2), atendidos por um VSB (Veículo de Suporte Básico). O SAMU-RP também faz a remoção de pacientes, caracterizada pelo transporte de pacientes entre hospitais, de casa para hospital ou vice-versa. O atendimento a ocorrências leves e remoção, faz com que a demanda pelo serviço aumente significativamente.

No SAMU-RP, não há um limite pré-estabelecido para o tamanho da fila de chamados. Os chamados em geral não são transferidos para outro sistema nos casos de congestionamento. As ambulâncias estão descentralizadas em cinco bases e o VSA não atende a chamados de prioridade moderada e leve. Por exemplo, no trabalho de Takeda (2000), as ambulâncias do SAMU-Campinas estavam todas centralizadas e os VSA's podiam atender a todos os tipos de chamados. As características do SAMU-RP são melhor descritas no Capítulo 3. O nível de serviço a ser oferecido e a configuração nestes sistemas devem ser escolhidos de acordo com os tipos de atendimentos mais solicitados, considerando, por exemplo, configurações geográficas do município, características da demanda local e regiões com diferentes índices de demanda ao longo do dia.

Na maioria dos estudos com o modelo hipercubo e com sistemas SAMU's encontrados na literatura, o SAE é dimensionado em apenas um período do dia (em geral, o período de pico). No entanto, durante um dia típico de operação de um SAMU, em determinados períodos pode haver diferenças significativas com relação ao número e à localização dos chamados. Além disso, conforme discussão adiante, alguns SAMU's, como o SAMU-RP, têm características de atendimento com prioridade na fila que requerem extensões do modelo hipercubo clássico que, até onde se tem conhecimento, nunca foram consideradas na literatura pesquisada.

Dessa maneira, uma pesquisa interessante para a análise de sistemas SAMU's é desenvolver uma abordagem baseada no modelo hipercubo que investigue o apoio à tomada de decisões, como localização e configuração das bases das ambulâncias do sistema, em termos das principais medidas de desempenho, para diferentes períodos do dia. Em cada período considerado, a configuração do sistema pode ser considerada de forma estática, ou seja, a configuração não muda dentro do período. Nesse contexto, é importante verificar a viabilidade e o benefício de se mudar a localização das ambulâncias ao longo do dia. Com essa análise, pode-se levar em conta ainda o *trade-*

off que existe entre o tempo médio de espera do sistema e a carga de trabalho das ambulâncias.

Diante disso, a fim de obter uma análise mais precisa do sistema, a abordagem a ser desenvolvida para sistemas SAMU, como o da cidade de Ribeirão Preto, deve incorporar políticas específicas de prioridade para as diferentes classes de usuários aguardando por atendimento na fila do sistema. Conforme mencionado, não temos conhecimento de trabalhos anteriores na literatura pesquisada que fizeram isso. Outro aspecto importante a ser considerado é a verificação do impacto no nível de serviço do sistema considerando o atendimento das remoções realizadas pelos VSB's.

1.1 Objetivos do trabalho

Dada a relevância dos SAMU's em cidades brasileiras, os objetivos principais desta tese são:

- (i) Propor uma abordagem baseada no modelo hipercubo de filas espacialmente distribuídas para localização das ambulâncias em mais de um período no dia de operação de um SAMU. Esses subperíodos podem ter diferentes características temporais e geográficas que podem ser consideradas no modelo. Assim, por meio desta abordagem, será possível analisar o sistema a partir das medidas de desempenho obtidas pelo modelo para cada subperíodo do dia e também analisar de forma independente cada período crítico do dia, possibilitando melhorar o atendimento em cada período estudado, do ponto de vista das medidas de desempenho internas e externas do sistema.
- (ii) Estender o modelo hipercubo clássico para tratar políticas específicas de prioridades na fila encontradas em certos SAMU's, como o da cidade de Ribeirão-Preto que atendem a diferentes classes de usuários. Quando a utilização do sistema é relativamente alta, de forma que a probabilidade de fila seja bem maior que zero, o tempo de resposta é o tempo médio de viagem mais o tempo médio de espera na fila. Assim, surge a necessidade de considerar prioridade na fila do modelo hipercubo para obter medidas de desempenho exatas em sistemas com essas características. Esta abordagem é especialmente importante para os usuários que apresentam risco de vida.

Outra questão importante que surge da análise do SAMU-RP é o

atendimento de remoções, discutido na Seção 2.1.2, visto que há um grande aumento do nível de utilização das ambulâncias. Isso faz com que o surgimento de filas de espera seja uma constante e sua administração torna-se um fator imprescindível. Por isso, em sistemas mais congestionados, torna-se importante analisar o impacto desta política de atendimentos nas medidas de desempenho do sistema.

Convém ressaltar que não se tem conhecimento de outros trabalhos na literatura consultada que trataram efetivamente desses objetivos descritos anteriormente em sistemas SAMU's. Para verificar a viabilidade da abordagem proposta em uma situação real, é realizado um estudo de caso no SAMU de Ribeirão Preto (SAMU-RP), assim como buscou-se avaliar o impacto do atendimento de remoções nas medidas de desempenho do sistema. Este trabalho pode ser visto como uma extensão das pesquisas em Takeda (2000) e Takeda *et. al* (2004, 2007), que estudaram a aderência do modelo hipercubo clássico e avaliaram a descentralização das ambulâncias no SAMU-Campinas. Convém salientar que, nesses estudos, não foram considerados diferentes períodos ao longo do dia de operação do SAMU-Campinas e também que as políticas de prioridade consideradas para atendimento dos chamados em fila foram as mesmas do modelo hipercubo clássico, ou seja, aproximações baseadas em matrizes de preferência fixa de despacho de ambulâncias e estratégias de divisão de átomos em camadas de subátomos (*layering*), conforme discutido em detalhes no Capítulo 4.

A abordagem proposta neste trabalho contribui para o processo de tomada de decisões em sistemas SAMU's, com respeito à quantidade e à localização das ambulâncias e ao nível de serviço oferecido pelo sistema. Para isso, considera as características aleatórias dos chamados em diferentes períodos do dia e as diferentes classes de usuários que, em um sistema com alta demanda, geram a necessidade de considerar, de forma precisa, a prioridade dos chamados na fila de espera.

1.2 Organização do texto

Esta tese está estruturada em seis capítulos.

O Capítulo 2 apresenta uma breve descrição dos SAE's e as características dos SAMU's em cidades brasileiras. São também discutidos em detalhes dois estudos de caso, o SAMU-Campinas, estudado em Takeda *et al.* (2004, 2007), e o Sistema de Atendimento Móvel de Urgência de Ribeirão Preto (SAMU-RP). Além disso, é feita

uma revisão dos principais modelos de localização descritivos e prescritivos relevantes para este trabalho.

O Capítulo 3 apresenta uma breve revisão sobre as principais definições e abordagens da teoria das filas com distribuições exponenciais que são utilizadas nesta tese, bem como o modelo hipercubo de filas espacialmente distribuídas. Devido a sua importância para este trabalho, uma revisão sobre modelos de filas com prioridades também é apresentada no final do capítulo.

O Capítulo 4 é dedicado à extensão do modelo hipercubo para considerar diferentes classes de usuários e disciplina de prioridade em uma fila finita, para a análise do SAMU-RP. Um pequeno exemplo do modelo proposto é apresentado, assim como suas medidas de desempenho, para ilustrar a discussão. É enfatizada a relação do modelo hipercubo proposto com o modelo $M/M/m/c$ e, ao final, é feito um procedimento para generalização do modelo proposto.

No Capítulo 5, inicialmente, é apresentada a descrição dos chamados do SAMU-RP e a validação das hipóteses do modelo hipercubo para este sistema. Em seguida, a abordagem proposta neste trabalho é aplicada para analisá-lo. Os resultados obtidos com a abordagem são analisados e comparados com o cenário original do SAMU-RP.

Finalmente, no Capítulo 6, são apresentadas as conclusões deste trabalho e são discutidas algumas perspectivas interessantes para pesquisas futuras.

2. Sistemas de Atendimento Emergencial e Modelos de Localização que Utilizam Teoria das Filas

Neste capítulo, são apresentadas as principais características de Sistemas de Atendimento Emergencial (SAE's). Particularmente, é feita uma descrição dos principais procedimentos dos SAMU's, com foco no SAMU da cidade de Ribeirão Preto. Também é apresentada uma revisão sobre os principais modelos de localização de facilidades, com ênfase em modelos que utilizam teoria das filas.

2.1 Sistemas de Atendimento Emergencial

Com o aumento da população urbana, torna-se necessário um esforço maior para garantir o acesso à saúde de todos os cidadãos. Em particular, os serviços de atendimento emergencial são importantes, pois garantem a qualidade de vida da população. A estimativa do IBGE é que a população do Brasil aumente de 183 milhões de habitantes em 2005 para 191 milhões de habitantes em 2009, o que aumenta a necessidade de sistemas eficazes e eficientes.

Nos últimos 50 anos, vem aumentando o interesse de pesquisadores em estudos de sistemas emergenciais em serviços urbanos (SIMPSON e HANCOCK, 2009), como corpo de bombeiros, patrulhamento policial e ambulâncias. Esses sistemas são, em geral, responsabilidade do setor público. Porém, empresas no setor privado, como empresas de segurança (monitoramento de alarmes) e sistemas de reparo da rede pública ou privada (água e esgoto, energia etc), também constituem serviços de atendimento emergenciais. Os SAE's são altamente complexos devido à grande incerteza com relação à distribuição espacial e temporal e à dependência da disponibilidade de seus servidores. Nesses sistemas, denominados sistemas servidores-para-clientes (*service-to-customer*), os funcionários devem se deslocar rapidamente até o local do chamado.

Em sistemas emergenciais, o tempo de resposta (tempo entre o chamado e a chegada do servidor no local da ocorrência) é uma medida de desempenho importante a ser considerada do ponto de vista do usuário. A carga de trabalho dos servidores

também é uma medida de desempenho importante, do ponto de vista do gerente do sistema. Essas duas medidas, em geral conflitantes porém altamente dependentes, devem ser gerenciadas juntamente com recursos limitados, a fim de o sistema oferecer um serviço eficiente e de qualidade para a população. Além do tempo de resposta, vários trabalhos, apresentados em Bodily (1978), Mendonça (1999) e Iannoni (2005) tratam a carga de trabalho dos servidores do sistema como um importante *trade-off* a ser analisado no sistema.

Savas (1969) e Takeda *et al.* (2004, 2007) mostraram como a descentralização das ambulâncias pode melhorar significativamente o tempo médio de resposta aos usuários. Taylor e Templeton (1980), Eaton *et al.* (1985) e Takeda *et al.* (2004, 2007) analisaram o fato de o sistema possuir diferentes tipos de ambulâncias, dependendo do grau de urgência do chamado. Todos esses trabalhos enfatizaram a importância do impacto da política de despacho dos servidores no sistema. De acordo com Chaiken e Larson (1972) e Swersey (1994), a política de despacho em SAE's pode ser definida como um conjunto de critérios que estabelecem:

- i) o número de servidores de cada tipo, em cada área geográfica, nos diferentes dias da semana;
- ii) a seleção de um servidor para atender a um chamado particular;
- iii) a determinação da localização de cada servidor;
- iv) a lista de preferência de despacho para cada área;
- v) redespacho ou realocação: sob quais circunstâncias as regras de despacho ou localização dos servidores podem ser alteradas.

Os estudos de Chaiken e Dormont (1978a), Chaiken e Dormont (1978b), Chelst (1978), Chelst (1981), Chelst e Barlach (1981), Green (1984), Green e Kolesar (1984a), Green e Kolesar (1984b) e Swersey (1994) analisaram políticas de despacho de viaturas de polícia do ponto de vista do número de viaturas enviadas para atender a um chamado. Um SAE deve ser eficiente e chegar de forma rápida até o local do acidente. Assim, é necessário que os SAE's mantenham informações atualizadas sobre a ocorrência, local do chamado, estado e local dos servidores. Nos SAE's em saúde, dependendo do tipo do chamado, é preciso transportar pessoal especializado (médicos e enfermeiras) e todo material (aparelhos e medicamentos) necessários para fazer o atendimento solicitado. Por exemplo, em um sistema de atendimento médico emergencial, quando ocorre um acidente de grandes proporções, é necessário enviar

uma UTI (Unidade de Terapia Intensiva) móvel, juntamente com médico e enfermeira ao local da ocorrência.

Outra particularidade dos SAE's é a cooperação entre os servidores. Se o servidor mais próximo estiver ocupado, outros servidores podem realizar o atendimento a um chamado em uma determinada área. Esses atendimentos são chamados de atendimentos *backup* e foram estudados em Iannoni (2005) e Iannoni *et al.* (2006b, 2008a, 2008b). No patrulhamento policial, em geral, as viaturas fazem a ronda em determinados setores da cidade e, na maior parte do tempo, estão se movendo em seu setor. As chamadas são recebidas por uma central e enviadas a uma viatura mais próxima do local da ocorrência. Larson (1971) estudou o impacto da política de despacho em carros de polícia por meio do modelo hipercubo, em termos das principais medidas de desempenho, como o tempo médio de viagem e a distância média percorrida.

Em Chelst (1975) há uma aplicação do modelo hipercubo em um sistema de patrulhamento policial. Nesse trabalho, foram avaliadas várias configurações do sistema que melhoram a carga de trabalho ou a distância percorrida nesse sistema, verificando-se que há um importante *trade-off* a ser analisado entre essas duas medidas de desempenho. Larson e Mcknew (1982) desenvolveram um método aproximado do modelo hipercubo, com a resolução de um sistema com N equações não-lineares. Em uma aplicação do modelo hipercubo no sistema de patrulhamento policial foram incorporados três estados dos servidores: livre, ocupado com um chamado ou ocupado em atividades de patrulha.

De acordo com Swersey (1994), as principais questões nesses estudos estão voltadas a solucionar problemas relacionados ao número de viaturas necessárias, determinação dos setores e atribuição de servidores, avaliação de desempenho desses sistemas e programação das equipes de atendimento. Esses problemas foram discutidos em Rider (1976), Ignall *et al.* (1982) e Swersey (1994). Ainda em Rider (1976), Swersey (1982) e Ignall *et al.* (1982) foram analisadas políticas de despacho de viaturas de bombeiro a fim de estudar o número de viaturas que devem ser enviadas quando ocorre um alarme. No trabalho de Oliveira (2003), foi avaliada a aplicação do modelo hipercubo de filas em um Centro de Emergência da Polícia Militar de Florianópolis - SC, verificando quais configurações alternativas do sistema (cenários alternativos)

melhorava o nível de serviço oferecido pelo sistema.

O Corpo de Bombeiros é um SAE em que seu serviço não é apenas relacionado ao combate a incêndios. Na maioria das cidades, os sistemas são equipados com carros que podem prestar serviços de primeiros socorros, podendo manter cooperação com as ambulâncias do serviço emergencial de saúde do município. Em Ignall *et al.* (1975), foi aplicado com sucesso o método da raiz quadrada para analisar o sistema do Corpo de Bombeiros de Nova Iorque em um projeto chamado *Rand Fire Project*. Em Swersey (1994), há uma descrição detalhada desse projeto. Esse método consiste em um método simples para determinar o número de servidores necessários para uma determinada região e baseia-se em estimar o tempo médio de viagem como função do número de unidades de atendimentos na região, sem a necessidade de um modelo de filas, como pode ser visto em Larson e Odoni (2007) e Swersey (1994). Em Kolesar e Blum (1973), foi mostrado que a distância média percorrida por viagem é inversamente proporcional à raiz quadrada do número de servidores da região por unidade de área. A constante de proporcionalidade depende da região e pode ser obtida por técnicas baseadas na modelagem por probabilidade geométrica ou simulação (LARSON e ODONI, 2007). A principal utilidade do método é descrever resultados obtidos para o tempo médio de viagem em diferentes políticas de alocação das ambulâncias (KOLESAR e BLUM, 1973).

No trabalho de Costa (2004) foi proposto um método para determinação de zonas de atendimento e localização para unidades de serviços emergenciais (ambulâncias), para que as áreas de atendimento fossem homogêneas segundo algum critério, como por exemplo: tempo médio de espera para o início do atendimento, tempo médio na fila de espera, desvio padrão dos tempos das unidades de trabalho por dia ou por atendimento. Foi feito um estudo de caso no sistema SIATE (Serviço Integrado de Atendimento ao Trauma em Emergência) da cidade de Curitiba-PR, sob a responsabilidade do Corpo de Bombeiros do Estado.

Desde o início da década de 70, a ambulância deixou de ser apenas veículo de transporte rápido de pacientes ao hospital mais próximo (SWERSEY, 1994 e TAKEDA, 2000). Muitas cidades possuem ambulâncias equipadas com suporte avançado de atendimento, composto por profissionais especializados, medicamentos e equipamentos que permitem o tratamento do paciente durante o transporte. Atualmente

muitas rodovias também possuem atendimento emergencial de saúde com ambulâncias bem equipadas, prestado por empresas privadas. Esse sistema tem a função de socorrer as vítimas de acidentes nas rodovias e, se necessário, fazer o transporte das mesmas ao hospital mais próximo. Há a característica de não admitir fila de espera. Se o servidor está ocupado, a chamada é transferida a um outro sistema como o corpo de bombeiros de uma cidade vizinha. Este sistema foi estudado por Mendonça (1999) e Iannoni (2005). Em algumas cidades brasileiras, existem sistemas de atendimento emergencial baseado no modelo francês, já comentado anteriormente, os SAMU's, que são objeto de estudo do presente trabalho e serão detalhados nas próximas seções.

2.1.1 Sistemas de Atendimento Móvel de Urgência (SAMU's)

De acordo com Takeda (2000) e Takeda *et al.* (2004, 2007), desde o início dos anos 90, vêm sendo feitos esforços importantes no sentido de melhorar a organização do atendimento médico-emergencial no Brasil. Conforme mencionado no Capítulo 1, os SAMU's são sistemas que operam de forma integrada com hospitais públicos em uma cidade ou região englobada por várias cidades. Atualmente, a rede SAMU conta com 147 serviços de atendimento móvel de urgência em todo o Brasil, por meio dos quais são atendidos 1273 municípios, totalizando a cobertura de 112 milhões de pessoas (Ministério da Saúde, 2009).

O sistema possui uma central com telefonistas que recebem os chamados e identificam a emergência. O chamado é transferido a um médico regulador que faz um diagnóstico da situação, orientando sobre as primeiras ações. Ao mesmo tempo, o médico regulador avalia o melhor procedimento para o paciente, orienta a pessoa a procurar um posto de saúde, ou realiza o despacho dos veículos de acordo com a prioridade do chamado e a disponibilidade dos recursos. O tempo total do atendimento emergencial realizado pelas ambulâncias engloba o tempo de recebimento e despacho das ambulâncias, um possível tempo de espera em fila, o tempo de viagem até o local da ocorrência, o tempo de atendimento em cena, o tempo de transporte ao hospital e o tempo de volta à base, no caso de esta não ser no hospital, pois as ambulâncias sempre retornam à base. Nos casos mais graves, o tempo de resposta está diretamente relacionado à sobrevivência e a alguma possível sequela aos pacientes.

Vários estudos nessa área foram feitos desde o final da década de 60. Os

primeiros modelos são relatados em Revelle *et al.* (1970) e Chaiken e Larson (1972). Desde então, muitas outras referências são encontradas nas últimas décadas em Kolesar e Swersey (1986), Revelle e Hogan (1989), Louveaux (1993), Swersey (1994), Owen e Daskin (1998), Chiyoshi *et al.* (2000) e Brotcorne *et al.* (2003). Os SAMU's estão instalados em cidades brasileiras de pequeno, médio e grande portes. O SAMU de São Paulo-SP, por exemplo, é um sistema de uma cidade de grande porte, possui suas ambulâncias descentralizadas e é caracterizado essencialmente por trabalhar com um grande número de ambulâncias. Esse problema foi recentemente estudado por Luque (2006).

Como descrito em Takeda (2000) e Takeda *et al.* (2004, 2007), o SAMU-Campinas passou por uma reorganização do seu sistema de atendimento de urgência em 1994. O município implantou o SAMU-192 com a função de ser o centro regulador das urgências médicas do município. No final dos anos 90, o sistema contava apenas com 10 ambulâncias, sendo que 8 delas eram veículos de suporte básico (VSB's) e 2 veículos de suporte avançado (VSA's). Todos os veículos permaneciam centralizados na base operacional do sistema, onde também se encontrava a central telefônica 192 para onde convergiam todos os chamados. O tamanho da fila era de, no máximo, a quantidade de ambulâncias disponíveis. Assim, a fila era limitada, em média, de 1 usuário por veículo em operação. A principal diferença entre os veículos estava na equipe técnica e nos equipamentos que os acompanhavam.

Os veículos de suporte básico (VSB's) possuíam equipamentos para o atendimento domiciliar de urgência. A equipe era composta por um enfermeiro e um auxiliar que se comunicavam com a base e realizavam atendimentos com até um certo grau de gravidade clínica e traumática, como é o caso de acidentes. Os veículos de suporte avançado (VSA's) eram mais aprimorados que os veículos básicos, eram uma UTI móvel e realizavam tratamentos mais complexos, não só de ordem clínica, mas também certos procedimentos cirúrgicos. A equipe acompanhante sempre era composta por um médico, um enfermeiro e um auxiliar de enfermagem.

A Figura 2.1 mostra a divisão da cidade de Campinas em átomos no final dos anos 90. O SAMU-Campinas definiu as regiões correspondentes às áreas de cobertura dos Centros de Saúde: Norte, Sul, Leste, Oeste e Central. Todas as ambulâncias estavam localizadas na região Central da cidade.

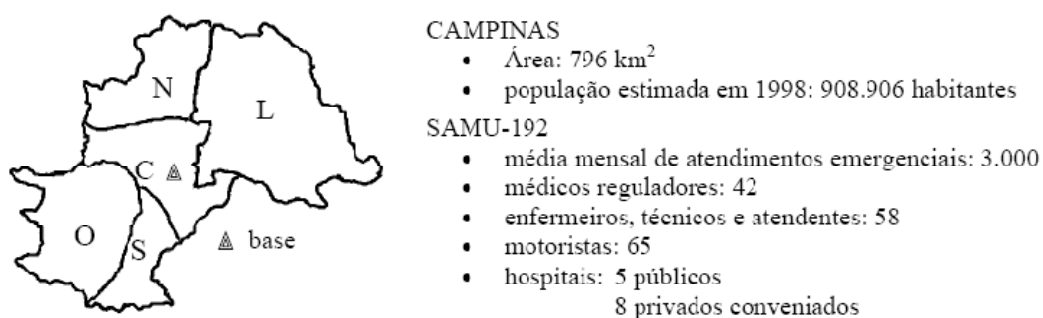


Figura 2.1 – Distribuição espacial do sistema SAMU-Campinas no final dos anos 90.
Fonte – TAKEDA (2000).

No estudo de Takeda (2000), foram consideradas duas classes de usuários: básica, definida pelos chamados atendidos por VSB's e avançada, definida pelos chamados atendidos por VSA's. Se houver um chamado básico, e todos os VSB's estiverem ocupados, os VSA's podem atendê-lo. Se houver um chamado avançado, e todos os VSA's estiverem ocupados, os VSB's podem atendê-lo. Dessa forma, não há *backup* parcial. Desde 1997, o SAMU-Campinas possui uma central de regulação com 8 linhas telefônicas, à disposição dos usuários pelo número 192, 24 horas por dia, por meio de 03 TARMs (Técnico Auxiliar de Regulação Médica) que registram o chamado. Essa central identifica o endereço e passa ao médico regulador para que este converse diretamente com o solicitante, avaliando a gravidade do caso e decidindo qual o recurso mais adequado à necessidade do solicitante. Todas as ligações eram passadas aos Médicos Reguladores (03 médicos reguladores de urgências clínicas e traumáticas e 01 médico regulador de urgências psiquiátricas). Todas as viaturas eram equipadas com GPS (*Global Positioning System*) que informam suas posições ao Operador de Frota.

Em 2008, o SAMU-Campinas estava operando com 15 ambulâncias (3 VSA's e 12 VSB's) e 1 ambulância de atendimento psiquiátrico, descentralizadas em Bases da Guarda Municipal, como segue: Base Cimcamp, 1 VSA e 1 VSB; Base Centro, 1 Viatura de Atendimento Psiquiátrica e 3 VSB's; Base GM Taquaral, 1 VSA e 3 VSB's; Base GM DIC 6, 1 VSA e 3 VSB's; Base GM Florence, 2 VSB's (PREFEITURA DE CAMPINAS, 2008). Em cidades como Ribeirão Preto (médio porte), atualmente as ambulâncias também estão descentralizadas. Assim como em Campinas, existem dois tipos de ambulâncias, as básicas e as avançadas, o sistema opera sem limitação de capacidade, podendo assim, operar sem limite para uma possível formação de fila. Há também SAMU's em cidades de pequeno porte, como por

exemplo, São Carlos, Araraquara e Jaú.

Neste estudo, em particular, foi escolhido o SAMU-RP por questões de conveniência de proximidade e acessibilidade aos dados. Os SAMU's possuem particularidades que os diferenciam de uma cidade para outra, como descrito anteriormente, que devem ser incorporadas ao modelo para a análise mais realista do sistema. Neste trabalho, é feito um estudo de caso detalhado no SAMU-RP.

2.1.2 O SAMU de Ribeirão Preto (SAMU-RP)

A população de Ribeirão Preto em 2005 estava estimada em aproximadamente 500 mil habitantes. Para 2008, estimou-se que a cidade tenha atingido 615.200 habitantes (IBGE, revisão 2008). De acordo com dados da TRANSERP (Empresa de Trânsito e Transporte Urbano de Ribeirão Preto), em 2005 ocorreram 13.859 acidentes de trânsito em Ribeirão Preto, ao passo que em 2008 ocorreram 15.796 acidentes. Assim, com o aumento do número de acidentes ao longo dos anos, há necessidade de que os sistemas de atendimento emergenciais urbanos sejam cada vez mais organizados e eficientes.

Além do SAMU, Ribeirão Preto também conta com outro serviço público para o atendimento da população, o Corpo de Bombeiros. A Tabela 2.1 mostra a distribuição do número de acidentes por tipo e dias da semana. Não há diferenças significativas de demanda quanto ao número de acidentes nos diferentes dias da semana.

Dia da semana	Acidentes com vítimas não pedestres	Acidentes sem vítimas	Atropelamentos	Total
Segunda-feira	432	1.950	34	2.416
Terça-feira	454	1.931	32	2.417
Quarta-feira	423	1.979	40	2.442
Quinta-feira	400	1.870	36	2.306
Sexta-feira	479	2.158	32	2.669
Sábado	456	1.705	30	2.191
Domingo	313	1.020	22	1.355
Total	2957	12.613	226	15.796

Tabela 2.1 – Quantidade de acidentes por tipo de dias da semana.
Fonte: TRANSERP – Prefeitura Municipal de Ribeirão Preto, 2008.

O SAMU-RP entrou em operação em 1996, criado a partir da iniciativa de profissionais da Secretaria Municipal da Saúde (LOPES, *et al.* 1999). Atualmente, tornou-se o principal responsável pelos atendimentos de urgência do município. O Corpo de Bombeiros também faz atendimentos de emergência (situação onde há risco iminente à vida) e urgência (situação onde não há risco iminente à vida) no município, porém é um sistema independente do SAMU-RP.

O SAMU-RP tem a responsabilidade de socorrer as vítimas e, nos casos mais urgentes, fazer ainda um pré-atendimento, no sentido de estabilizar a vítima e depois encaminhá-la a um hospital próximo. Atualmente, uma Central de Regulação Médica se encontra no prédio da Secretaria da Saúde de Ribeirão Preto no centro da cidade e não possui veículos. Onde são recebidos os chamados pelo número 192 e encaminhados para o médico regulador, que julga e decide sobre a gravidade do caso, orientando o atendimento a ser realizado e disponibilizando o envio de recursos.



Figura 2.2 – Ambulâncias VSB's do SAMU-RP.

O sistema conta com 1 veículo de suporte avançado e 9 veículos de suporte básicos, que são despachados, dependendo da gravidade do chamado e da sua localização. Os veículos estão distribuídos em cinco bases de atendimento no município, situadas em hospitais ou postos de saúde. Os veículos VSA e VSB's do SAMU-RP se diferem basicamente com relação aos equipamentos e a equipe enviada ao chamado. Os VSB's (um exemplo pode ser visto na Figura 2.2) são constituídos por uma equipe formada por um auxiliar de enfermagem e um motorista. Na Figura 2.3, é apresentada a parte interna do VSB e os equipamentos de primeiros socorros, como

imobilizadores, oxigênio e medicamentos básicos. Os VSB's são enviados preferencialmente para atender casos urgentes, mas também atendem emergências quando o VSA está ocupado.



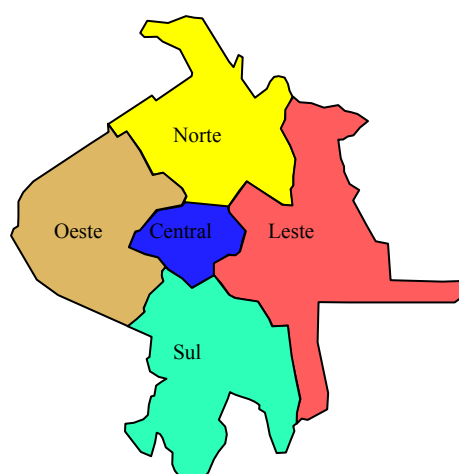
Figura 2.3 – Equipamentos dos veículos VSB's do SAMU-RP.

Em sua maioria, as chamadas atendidas pelos VSB's são urgências e são divididas em duas classes: moderada, como fratura exposta, bronquite, quadro pneumônico, choque cardiogênico entre outros; e leve, situações em que a o paciente apresenta um quadro menos preocupante como, lombalgia, febre, dor de cabeça entre outros. Por outro lado, o VSA é constituído por uma equipe formada por um médico, um enfermeiro e um motorista, e os equipamentos são mais sofisticados que os do VSB (Figura 2.4), a fim de que haja melhor atendimento e seja possível estabilizar o paciente até a sua chegada ao hospital. O VSA é enviado a um chamado apenas quando há uma emergência, como: comprometimento da via aérea, comprometimento da dinâmica respiratória, comprometimento da dinâmica circulatória, comprometimento da função neurológica ou comprometimento funcional de extremidades.

A cidade de Ribeirão Preto está dividida em cinco regiões: Norte, Sul, Leste, Oeste e Central (Figura 2.5). Essa é a divisão utilizada pelo SAMU-RP e a distribuição das ambulâncias é descentralizada, localizada em cinco postos de saúde. A Região Central possui três ambulâncias, sendo duas delas VSB's e a outra VSA. A Região Norte, Sul e a Região Oeste possuem duas VSB's cada e a Região Leste possui uma VSB.



Figura 2.4 – Equipamentos do VSA no SAMU-RP.



RIBEIRÃO PRETO

Área: 652 km²

População: 547.417 habitantes

SAMU - 192

Média mensal de atendimentos emergenciais: 10000 chamados, sendo 4100 chamados de alta, média e baixa complexidade. O sistema realiza 5900 atendimentos de remoção.

Médicos reguladores: 20

Enfermeiros e técnicos: 75

Motoristas: 68

Hospitais: 1 público

2 filantrópicos

Figura 2.5 – Distribuição espacial do sistema SAMU - Ribeirão Preto em 2005.

Uma característica importante do SAMU-RP é o atendimento de três classes de usuários: alta, média e baixa complexidade. Além disso, realizam a remoção agendada de pacientes que se caracteriza pelo transporte de pacientes entre hospitais ou de casa para hospital. Esse tipo de atendimento faz com que haja um aumento significativo da carga de trabalho das ambulâncias, resultando na formação de maiores filas de espera (um chamado dificilmente é enviado a outro sistema). O sistema usa política de despacho considerando prioridade de atendimento na fila; assim, são atendidos os usuários na ordem: alta, média e baixa complexidade, respectivamente, e com prioridade mais baixa para as remoções. Como visto adiante, para fins de simplicidade de modelagem e análise, foi considerado que as remoções e os chamados de baixa complexidade formam uma única classe de usuários.

2.2 Modelos de localização

Na literatura pesquisada, encontram-se diversos modelos de localização descritivos e prescritivos. Os modelos descritivos ajudam os gerentes dos sistemas a comparar cenários, calculando medidas de desempenho para cada um. Enquanto que os prescritivos são modelos de otimização, nos quais uma ou mais medidas de desempenho são foco de uma função objetivo (critério a ser otimizado). Os modelos descritivos de localização probabilística são utilizados para análise de sistemas a fim de descrever suas principais medidas de desempenho. Eles não apontam uma configuração ideal de operação do ponto de vista de uma medida de desempenho (como os modelos prescritivos). A principal importância dos modelos de localização probabilística é que eles consideram as características aleatórias do sistema, como por exemplo, incertezas com relação a taxas de chegada dos usuários, a taxas de serviço ou ainda, a localização dos servidores e tempos de viagem.

Em problemas de localização de SAE's, busca-se prover a cobertura das áreas de demanda. Uma área é considerada coberta se ela está a menos de uma distância (tempo) crítica de pelo menos um dos servidores existentes, em geral independentemente de eles estarem disponíveis ou não. Os problemas de localização baseados em cobertura de conjuntos são, em sua maioria, prescritivos e seu objetivo é encontrar a localização de unidades servidoras, de forma a satisfazer alguma medida de desempenho. O modelo de localização mais simples é o Problema de Localização para Cobertura de Conjuntos (PLCC). Um problema relacionado a ele é a localização de p servidores, de tal forma que a distância máxima de qualquer área de demanda ao servidor mais próximo seja a mínima possível. No Problema de Localização de Máxima Cobertura (PLMC), o objetivo é localizar um número pré-estabelecido de servidores p , compatível com os recursos disponíveis, de tal forma que, a máxima população possível de uma dada região seja coberta a menos de uma distância crítica S pré-definida (CHURCH e REVELLE, 1974). Daskin (1983) desenvolveu uma heurística de substituição de vértices para a resolução do problema. Por meio dos estudos de Revelle e Hogan (1989), foram desenvolvidos modelos de cobertura adicional, analisando a possibilidade de os sistemas estarem congestionados.

Alsalloum e Rand (2006) propuseram uma variação do PLMC, quando consideraram a probabilidade de cobertura de uma população dentro de um tempo

especificado, com o objetivo de localizar as bases de EMS e, em seguida, determinar o número mínimo de veículos de forma a garantir um nível de serviço. Asseguram ainda, que qualquer demanda localizada dentro da área de serviço de uma base poderia encontrar pelo menos um veículo disponível. Esse estudo foi aplicado na cidade de Riyadh, Arábia Saudita. Em Revelle e Hogan (1989), foi estudado o Problema de Localização de Máxima Disponibilidade (PLMD), que busca localizar p servidores tal que um máximo número de chamadas a um serviço de emergência tenha um servidor disponível a menos de uma distância crítica S com confiabilidade α . Savas (1969) utilizou simulação computacional para analisar o sistema de despacho de ambulâncias de um distrito do Brooklyn, em Nova Iorque. Esse autor analisou também custos e benefícios em adicionar ambulâncias no sistema.

Um avanço no desenvolvimento da teoria relacionada à localização de servidores de emergência foram os modelos de cobertura probabilísticos que consideram a aleatoriedade na disponibilidade dos servidores um fator importante em modelos de cobertura. A principal importância desses modelos é que eles consideram a aleatoriedade (incertezas) do sistema, por exemplo: a probabilidade de ocupação dos servidores é uma medida essencial a ser considerada, a natureza aleatória dos chamados, a incerteza quanto ao local do chamado, horário e duração dos atendimentos. Esses são fatores que aumentam a complexidade do sistema de atendimento de emergência, como os SAMU's. Eles não apontam a configuração ótima com relação à localização ou à política de despacho dos servidores para o sistema. Nos trabalhos de Owen e Daskin (1998), Chiyoshi *et al.* (2000), Marianov e Serra (1998, 2001a, 2001b, 2003), Marianov e Ríos (2000), Boffey *et al.* (2007), Galvão e Morabito (2008) e Corrêa (2008) podem ser observados detalhes sobre esses e outros trabalhos nessa área.

2.2.1 Modelos que utilizam Teoria das filas

Os modelos descritivos de localização probabilística foram desenvolvidos para análise e planejamento de sistemas de atendimento emergencial. Em especial, o modelo hipercubo proposto por Larson (1974) (baseado em filas espacialmente distribuídas) trata de sistemas urbanos em que os servidores se deslocam para fornecer algum tipo de serviço para os clientes (*server-to-customer*). Esse modelo é bastante

utilizado em problemas de localização probabilística como em Batta *et al.* (1989), Saydam *et al.* (1994), Chiyoshi *et al.* (2003), Galvão *et al.* (2003), Galvão *et al.* (2005), Iannoni (2005), entre outros. Há vários trabalhos que utilizam modelos descritivos e probabilísticos aplicados a SAE's. Alguns exemplos são citados a seguir. Alguns desses trabalhos envolvem o estudo da aplicação do modelo hipercubo em problemas de localização probabilística.

Bell e Allen (1969) utilizaram o modelo $M/G/\infty$ para representar um sistema emergencial de saúde. Por não ser um modelo de filas espacialmente distribuídas, esses autores consideraram os servidores centralizados, localizados juntamente com a central de chamados. Taylor e Templeton (1980) desenvolveram uma aplicação para a região central da cidade de Toronto. O modelo estabelece duas classes diferentes de usuários: i) pacientes que precisam de atendimento emergencial imediato, que correm risco de vida; ii) pacientes que podem esperar em fila em situações que não correm risco de vida. As principais medidas de desempenho calculadas foram a probabilidade de uma emergência encontrar um determinado número de ambulâncias disponíveis, o tempo médio de espera para uma urgência e a probabilidade do tempo de espera de uma urgência ser superior a um valor considerado aceitável. Dessa forma, foi possível determinar o número de ambulâncias necessárias para realizar os atendimentos, de forma a garantir o nível de serviço desejado.

Batta *et al.* (1989) combinam o modelo hipercubo e uma heurística de substituição de vértices com o objetivo de determinar a localização dos servidores, de tal forma a maximizar a cobertura esperada. A heurística de substituição de vértices foi desenvolvida originalmente por Teitz e Bart (1968) para o problema das p -medianas. Esse procedimento é realizado a cada configuração encontrada, buscando encontrar configurações que melhorem o valor da cobertura esperada. Em Swersey (1994) há uma análise detalhada dos métodos aplicados até o início dos anos 90 para solucionar problemas de localização probabilística em sistemas emergenciais.

Em Mendonça e Morabito (2000), foi avaliada a aplicação do modelo hipercubo no sistema Anjos do Asfalto na rodovia Presidente Dutra - BR, obtendo as medidas de desempenho do sistema e avaliando cenários alternativos a fim de melhorar o balanceamento do *workload* das ambulâncias. Chiyoshi *et al.* (2000) analisou a integração do modelo hipercubo à modelagem probabilística de localização de

servidores, em sistemas em que a aleatoriedade na disponibilidade das ambulâncias deve ser um fator a ser considerado. Em Chiyoshi *et al.* (2001) e Morabito *et al.* (2007), foi analisada a utilização do modelo hipercubo no caso de haverem servidores não homogêneos no sistema, como é o caso de vários SAMU's em algumas cidades do Brasil. Os mesmos autores ainda estudaram métodos de solução desses sistemas para cada caso. O método de Gauss-Siedel é um dos mais indicados para o caso da análise de sistemas emergenciais com servidores não-homogêneos.

Em Figueiredo *et al.* (2005), foram estabelecidos átomos geográficos para a utilização do Sistema Resgate Saúde da prefeitura municipal da cidade de São José dos Campos-SP, juntamente com o Corpo de Bombeiros, a fim de equilibrar a distribuição populacional nas regiões da cidade. Utilizou-se o modelo hipercubo de filas e encontraram-se as medidas de desempenho do sistema. O trabalho de Takeda *et al.* (2004, 2007) avaliou a utilização do modelo hipercubo no SAMU-Campinas, mostrando que as medidas de desempenho calculadas pelo modelo hipercubo são precisas. Avaliaram cenários alternativos e estudaram alternativas de descentralização das ambulâncias a fim de melhorar algumas medidas de desempenho, como tempo de viagem e *workload* das ambulâncias.

Atkinson *et al.* (2006, 2008) tratam das diferenças nos tempos de viagem da primeira e segunda bases preferenciais de ambulâncias para atender os trechos (átomos) em um SAE de uma rodovia. A taxa de serviço pode ser significativamente diferente nos dois casos. O modelo assume que uma ambulância pode estar em três estados: (1) livre; (2) ocupado, atendendo um paciente de primeira preferência; (3) ocupado, atendendo um paciente de segunda preferência. A fim de calcular medidas de interesse, como a probabilidade de perda (p_{loss}), é necessário resolver um sistema de 3^m equações lineares, em que m é o número de bases. O modelo estudado admite que as taxas de serviço possam variar dependendo do tipo do chamado. A intratabilidade desse modelo significa que soluções exatas para a p_{loss} podem ser obtidas somente para poucas bases (valores pequenos de m). Foram desenvolvidas duas heurísticas para resolver o modelo hipercubo modificado, considerando três estados para cada servidor.

Singer e Donoso (2008) usaram simples expressões matemáticas para avaliar um serviço de ambulância no setor privado em Santiago no Chile. O serviço de ambulância foi descrito em termos dos principais parâmetros de operação e variáveis de

decisões estratégicas. Com o auxílio da teoria das filas, foram calculados os indicadores de desempenho chave do ponto de vista do administrador e do paciente. Foi avaliado se o desempenho histórico é consistente com os recursos disponíveis e estimado o impacto de algumas mudanças operacionais, como redução do tempo de ciclo ou aumento da frota. A cobertura geográfica das bases foi otimizada com o objetivo de minimizar o tempo médio de viagem.

2.2.2 Modelos que combinam o modelo hipercubo com procedimentos de otimização

Conforme mencionado, o modelo hipercubo possibilita a utilização de filas espacialmente distribuídas em modelos de localização probabilísticos. É um modelo descritivo que possibilita a análise de vários cenários, porém, se combinado com técnicas de otimização e heurísticas, pode ser utilizado para resolver problemas de localização probabilísticos. Alguns trabalhos enfocam a utilização do modelo hipercubo juntamente com um procedimento de otimização, a fim de melhorar o sistema do ponto de vista de alguma(s) medida(s) de desempenho. Dentre os modelos, podem-se citar Iannoni (2005) e Iannoni *et al.* (2008a, 2008b) que estudaram a combinação do modelo hipercubo com um algoritmo genético para otimizar a configuração e operação de SAE's em rodovias. O trabalho apoiou decisões relacionadas ao planejamento e operação desses sistemas, como por exemplo, determinar o tamanho da área de cobertura para cada ambulância, de forma a minimizar o tempo médio de resposta ao usuário e/ou o desbalanceamento da carga de trabalho das ambulâncias. Consideraram, ainda, particularidades dos SAE's em rodovias, desenvolvendo extensões para o modelo hipercubo de filas, como *backup* parcial (os servidores podem não atender a todos os átomos) e duplo despacho de ambulâncias (dois servidores são enviados a um mesmo chamado).

Geroliminis *et al.* (2006) propuseram um modelo de localização que minimiza o tempo médio de resposta utilizando uma generalização do modelo hipercubo. Os resultados foram comparados com outros modelos de localização, como o MCLP (CHURCH e REVELLE, 1974) e *p*-mediana (HAKIMI, 1964), e mostraram que o modelo é uma importante ferramenta de otimização, principalmente em sistemas com

alta demanda, quando os servidores podem não estar disponíveis quando surgir um chamado. Além dos estudos acima, outros estudos combinaram procedimentos de otimização com o modelo hipercubo, como por exemplo, Batta *et al.* (1989), Saydam e Aytug (2003), Chiyoshi *et al.* (2003), Galvão *et al.* (2005), Rajagopalan *et al.* (2008), Erkut *et al.* (2008) e Ingolfsson *et al.* (2008).

2.2.3 Modelos de localização dinâmica

A localização de facilidades indistinguíveis e sujeitas a congestão foi estudada em Marianov e Ríos (2000), Marianov e Serra (1998, 2003) e Boffey *et al.* (2007). Esses estudos modelaram sistemas utilizando o modelo de filas $M/M/m/\infty$. Alguns trabalhos têm sido direcionados na utilização de localização dinâmica de ambulâncias a fim de melhorar o desempenho de SAE's ao longo do tempo. Gendreau *et al.* (2006) formularam e resolveram um problema dinâmico que surge com a re-localização dos veículos médicos. O objetivo era maximizar a cobertura esperada no tempo, sujeito a um limite do número de bases que podiam mudar a cada evento. Foi formulado um problema de programação linear inteira chamado de MECRP (Problema de Re-localização de Máxima Cobertura Esperada). Utilizaram um modelo de otimização, na qual várias soluções são computadas em antecipação dos eventos futuros e a solução, se disponível, é implementada quando um evento ocorre. Esse método foi aplicado com sucesso no problema de relocalização de ambulâncias na cidade de Montreal no Canadá.

No trabalho de Rajagopalan *et al.* (2008), foi formulado um modelo de localização dinâmica de cobertura disponível (DACL), que determina o número mínimo de ambulâncias e suas localizações para cada período de tempo, em cada mudança significativa no padrão de demanda. Um importante aspecto do DACL é a incorporação do modelo hipercubo, relaxando a suposição de que todos os servidores têm a mesma probabilidade de estarem ocupados e operarem independentemente. Esse modelo, que estende o modelo Q-PLSCP de Marianov e Revelle (1996) para múltiplos períodos, incorpora a probabilidade de a ambulância estar ocupada usando o algoritmo aproximado do modelo hipercubo de Jarvis (1985). Eles desenvolveram um algoritmo de busca tabu para resolver o modelo. Os resultados mostraram que o modelo

desenvolvido produz alta qualidade das soluções em um tempo computacional razoável.

Ingolfsson *et. al.* (2008) mostraram que o tempo de viagem entre pares de pontos são significantes e altamente variáveis. Apresentaram um modelo de cobertura mais realista para alocar um número específico de ambulâncias em suas bases, que depende da disponibilidade das ambulâncias, dos atrasos e dos tempos de viagem. Foi feito um estudo de caso na cidade de Alberta, Canadá, em que foi mostrado que a inclusão de variabilidades nos atrasos tem um impacto substancial na solução desses modelos. Corrêa (2008) discutiu e propôs métodos de solução para dois problemas de localização de facilidades: o problema de localização de facilidades não capacitados (UFLP) e o problema probabilístico de localização-alocação de máxima cobertura (PPLAMC). Esse autor usou técnicas de modelagem nas restrições do problema por meio de um grafo, particionando-o em *clusters* (agrupamento de vértices bem definidos). Para resolver o problema (PPLAMC) foram utilizados um algoritmo genético construtivo (AGC) e uma busca de agrupamentos (*clustering search* - CS).

2.2.4 Modelos de simulação

Os modelos de simulação em geral permitem estudar sistemas com maior nível de detalhes que os modelos analíticos, que geralmente exigem muitas hipóteses simplificadoras na sua construção. Alguns sistemas requerem um grande número de simplificações para serem modelados de forma analítica, comprometendo sua análise. Nesses casos, a simulação pode ser uma ferramenta mais adequada para analisar tais sistemas. Por exemplo, nos trabalhos de Pegden *et. al.* (1995) e Banks (1998) são discutidas as principais vantagens do uso da simulação e os aspectos a serem considerados para a análise adequada de um sistema por meio de um modelo de simulação. A abordagem permite analisar e testar configurações alternativas facilmente, como diferentes políticas de operação, aumento ou redução nos recursos disponíveis, aumento da demanda e muitas outras modificações. As medidas de desempenho dos vários cenários analisados são comparadas de forma a identificar o cenário mais promissor.

Conforme mencionado, Savas (1969) foi um dos pioneiros a utilizar um modelo de simulação para analisar um sistema de ambulâncias localizado em Brooklyn,

considerando características aleatórias. No modelo original, todas as ambulâncias estavam localizadas em um hospital e os resultados obtidos com a simulação mostraram que o tempo de resposta a um chamado pode ser reduzido em até 10%. Alguns cenários foram avaliados localizando algumas ambulâncias em garagens próximas das regiões com maior número de ocorrências. Também foram avaliados custos e benefícios de adicionar ambulâncias no sistema. As melhores alternativas foram as que consideraram adicionar e dispersar ambulâncias em garagens satélites.

Além da possibilidade da análise de um sistema por meio de modelos de simulação, outro caminho que vem sendo largamente usado em pesquisa operacional é a utilização da simulação como instrumento de validação dos modelos analíticos. Por meio da simulação pode-se verificar se as simplificações adotadas nesses modelos não comprometem os resultados da análise, pois os resultados das duas formas de análise devem convergir, a não ser pela diferença devida aos erros amostrais da simulação. Há vários estudos utilizando simulação em sistemas de atendimento emergencial para validar modelos analíticos. Ignall *et. al.* (1978) utilizaram simulação para validar simples modelos analíticos aplicados aos sistemas de emergência, tais como: o modelo de fila $M/M/m$ na análise de sistema de patrulhamento policial, o modelo da raiz quadrada utilizado por Kolesar e Blum (1973) para estimar a média de resposta das viaturas de bombeiro, entre outros. O estudo mostrou que esses modelos podem ser seguramente utilizados para análise dos sistemas reais, de forma mais econômica que a simulação.

Iannoni (2005) utilizou um modelo de simulação para avaliar se as medidas de desempenho obtidas das modificações realizadas no modelo hipercubo para considerar *backup* parcial estavam bastante próximas, a não ser pelo erro amostral. A simulação também foi utilizada para validação de modelos analíticos envolvendo SAE's em Fitzsimmons (1973), Kolesar e Blum (1973), Ignall *et. al.* (1978), Swersey (1982), Goldberg *et. al.* (1990), Iannoni (2005) e Iannoni *et. al.* (2008a, 2008b, 2009). Ignall *et. al.* (1978) e Larson e Odoni (2007) enfatizam que se comprovado que um modelo analítico pode ser usado em um sistema real, ele deve ser utilizado para análises futuras. Em geral, a simulação é um método computacionalmente mais caro e a interpretação dos resultados é relativamente mais difícil que nos métodos analíticos. Ignall *et. al.* (1978) discutem as vantagens da utilização dos métodos analíticos ao invés da simulação. Uma vantagem dos métodos analíticos é que eles podem ser incorporados

em outros modelos e exigem menor detalhamento e análise dos dados de entrada, a um custo menor do que a simulação. Um modelo de simulação pode ser construído para verificar o quanto as simplificações de um modelo analítico comprometem os resultados.

Zaki e Cheng (1997) construíram um modelo de simulação para analisar o sistema de patrulhamento policial em Richmond, no estado de Virgínia nos Estados Unidos. Consideraram diferentes formas alternativas de alocação de veículos e várias complexidades do sistema, como zonas não-homogêneas de demanda cujo processo de chegada nem sempre é exponencial, variação das condições e operação do sistema de acordo com o período com o dia da semana, estação do ano, horas de pico e outros fatores. Por meio de testes de simulação, Berman e Vasudeva (2005) analisaram o cálculo de medidas de desempenho aproximadas, como tempo médio de resposta, em SAE's para serviços do setor público. Consideraram ainda, em sua aproximação, que os tempos de serviço são independentes e identicamente distribuídos. No próximo capítulo, é feita uma breve revisão sobre os modelos de filas com distribuições exponenciais e o modelo hipercubo em sua forma original, que são utilizados nesta tese.

3. Teoria das Filas com Distribuições Exponenciais e Modelo Hipercubo

Neste capítulo, é apresentada uma breve revisão sobre as principais definições e abordagens da teoria das filas com distribuições exponenciais, com o objetivo de mostrar as relações entre os modelos que são utilizados nesta tese. É dada ênfase na descrição do modelo hipercubo em sua forma original e sua relação com o caso em que o processo de chegada é Poisson e há m servidores idênticos com tempos de serviço exponencialmente distribuídos. No final do capítulo, procede-se a uma revisão do modelo que diferencia os usuários em classes de acordo com o tipo de serviço solicitado, em que a chegada dos usuários em cada classe é Poisson, há m servidores idênticos com serviço exponencialmente distribuídos e utiliza-se disciplina de prioridade na fila.

3.1 O processo de Poisson e a distribuição exponencial

O processo de Poisson descreve o número de eventos $(N(t))$ que ocorrem em um intervalo de tempo. Segundo Feller (1968), o processo começa no instante 0 a partir do estado $n = 0$ (sistema vazio) e é definido por dois postulados:

1. As transições a partir de um estado n são possíveis apenas para $n + 1$.
2. Qualquer que seja o estado n em um instante t , a probabilidade de ocorrer um evento dentro de um pequeno intervalo de tempo entre t e $t + \Delta t$ é igual a $\lambda \Delta t$ (λ é a taxa de chegada de usuários). A probabilidade de ocorrer mais que um evento nesse intervalo é $o(\Delta t)$,

$$\text{tal que } \lim_{\Delta t \rightarrow 0} \frac{o(\Delta t)}{\Delta t} = 0.$$

Considere que o sistema esteja no estado n , no instante t , e não ocorra nenhum evento entre t e $t + \Delta t$. Então, o sistema continuará no estado n com probabilidade $P_n(t)[1 - \lambda \Delta t] + o(\Delta t)$. Caso o sistema esteja no estado $n - 1$, no

instante t , e ocorra um evento entre t e $t + \Delta t$, o sistema muda para o estado n no instante $t + \Delta t$, com probabilidade $P_{n-1}(t)\lambda\Delta t$. Assim, no instante $t + \Delta t$ o sistema estará no estado n com probabilidade $P_n(t + \Delta t)$, expressa pela Equação (3.1).

$$P_n(t + \Delta t) = P_n(t)[1 - \lambda\Delta t] + P_{n-1}(t)\lambda\Delta t + o(\Delta t). \quad (3.1)$$

Observe que para o caso particular de $n = 0$, no instante $t + \Delta t$, a equação não possui o termo $P_{n-1}(t)\lambda\Delta t$, pois não existe o estado $n - 1$ que antecede $n = 0$. Assim, a probabilidade de o sistema estar no estado $n = 0$ no instante $t + \Delta t$ ($P_0(t + \Delta t)$) é dada pela Equação (3.2).

$$P_0(t + \Delta t) = P_0(t)[1 - \lambda\Delta t] + o(\Delta t). \quad (3.2)$$

As equações (3.1) e (3.2) podem ser reescritas na forma:

$$\begin{aligned} \frac{P_n(t + \Delta t) - P_n(t)}{\Delta t} &= -\lambda P_n(t) + \lambda P_{n-1}(t) + \frac{o(\Delta t)}{\Delta t}, \\ \frac{P_0(t + \Delta t) - P_0(t)}{\Delta t} &= -\lambda P_0(t) + \frac{o(\Delta t)}{\Delta t}. \end{aligned}$$

No limite de $\Delta t \rightarrow 0$, temos que:

$$\begin{aligned} \frac{dP_0(t)}{dt} &= -\lambda P_0(t), & \text{para } \mathbf{n} = \mathbf{0}. \\ \frac{dP_n(t)}{dt} &= -\lambda P_n(t) + \lambda P_{n-1}(t), & \text{para } \mathbf{n} \geq \mathbf{1}. \end{aligned} \quad (3.3)$$

O sistema (3.3), de infinitas equações, mostra uma relação de recursão tal que, quando $n = 0$ temos $P_0(t) = e^{-\lambda t}$. Substituindo esse resultado para a equação de $n = 1$, obtém: $\frac{dP_1(t)}{dt} = -\lambda P_1(t) + \lambda e^{-\lambda t}$, resultando em: $P_1(t) = (\lambda t) e^{-\lambda t}$. Assim por diante, obtendo por indução:

$$P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, \text{ para } n = 0, 1, 2, \dots \quad (3.4)$$

A Equação (3.4) corresponde à distribuição de Poisson, obtida a partir dos postulados 1 e 2. Pode-se notar que a probabilidade de não ocorrer nenhuma chegada ($n=0$) até o instante t é $P_0(t) = e^{-\lambda t}$, que é exatamente a probabilidade de que a primeira chegada ocorra depois do instante t em uma distribuição exponencial ($P(T > t)$).

$$P_0(t) = P(T > t) = e^{-\lambda t} \quad (3.5)$$

A Equação (3.5) mostra uma importante relação entre o processo de Poisson e a distribuição exponencial, que pode ser enunciada da seguinte forma: se o número de vezes que o evento ocorre ao longo do tempo possuir uma distribuição de Poisson (Equação (3.4)), os intervalos de tempo entre as ocorrências consecutivas dos eventos seguem uma distribuição exponencial. Essa relação entre a distribuição exponencial e Poisson é utilizada para a construção das equações de equilíbrio das filas Markovianas e, no Capítulo 5, para verificar a hipótese de que os chamados do SAMU-RP chegam a partir de um processo de Poisson.

Em uma distribuição exponencial, a probabilidade de uma variável aleatória contínua T ser maior que $t + s$, dado que T é maior que s não depende de s , é definida da seguinte forma:

$$P(T > s + t / T > s) = \frac{P(T > s + t \cap T > s)}{P(T > s)}. \quad (3.6)$$

Como $P(T > s + t \cap T > s) = P(T > s + t)$, então:

$$P(T > s + t / T > s) = \frac{P(T > s + t)}{P(T > s)}.$$

$$P(T > s + t / T > s) = \frac{e^{-(s+t)\lambda}}{e^{-s\lambda}} = e^{-t\lambda} = P(T > t). \quad (3.7)$$

Esta é uma propriedade da distribuição exponencial denominada “falta de memória”, na qual também é uma propriedade das filas Markovianas. A distribuição exponencial é utilizada na Seção 3.5 e nos Capítulos 4 e 5 para determinar a probabilidade de um servidor terminar o serviço antes dos demais, caso todos estejam ocupados, ou ainda, a probabilidade de um chamado surgir primeiro no átomo j (desde que os chamados cheguem de acordo com o processo de Poisson).

Tomando-se um exemplo ilustrativo: um sistema com três servidores (A , B e C) e seus tempos de serviço considerados variáveis aleatórias X , Y e Z independentes e exponencialmente distribuídos com parâmetros λ_x , λ_y e λ_z , respectivamente. A probabilidade de o servidor A ficar livre primeiro é a probabilidade de o servidor A terminar o serviço com menor tempo entre os três servidores ($P[X = \min(X, Y, Z)]$), e

equivale dizer que o servidor A termina o serviço com tempo menor do que o mínimo dos tempos de serviço dos servidores B e C ($P[X < \min(Y, Z)]$). Portanto, é necessário saber qual é a distribuição relacionada a uma variável aleatória W , definida como o mínimo de duas variáveis aleatórias exponenciais independentes, Y e Z ($W = \min(Y, Z)$), com parâmetros λ_y e λ_z .

$$P(W > w) = P(Y > w \text{ e } Z > w),$$

$$P(W > w) = P(Y > w)P(Z > w),$$

$$P(W > w) = \int_w^\infty \lambda_y e^{-\lambda_y y} \int_w^\infty \lambda_z e^{-\lambda_z z} dy dz$$

$$P(W > w) = e^{-\lambda_y w} e^{-\lambda_z w} = e^{-(\lambda_y + \lambda_z)w}.$$

Nessas condições, W é uma exponencial com parâmetro $\lambda_w = \lambda_y + \lambda_z$. A questão do exemplo ilustrativo é encontrar a probabilidade de o servidor A terminar o serviço antes dos demais ($P(X \leq w)$), da seguinte forma:

$$\begin{aligned} P(X \leq w) &= \int_{X=0}^\infty \int_{W=X}^\infty \lambda_x e^{-\lambda_x x} \lambda_w e^{-\lambda_w w} dx dw, \\ &= \int_{X=0}^\infty \lambda_x e^{-\lambda_x x} \int_{W=X}^\infty \lambda_w e^{-\lambda_w w} dx dw, \\ &= \int_{x=0}^\infty \lambda_x e^{-(\lambda_x + \lambda_w)x} dx, \\ &= \frac{\lambda_x}{\lambda_x + \lambda_w} \int_{x=0}^\infty (\lambda_x + \lambda_w) e^{-(\lambda_x + \lambda_w)x} dx, \\ P(X \leq w) &= \frac{\lambda_x}{\lambda_x + \lambda_y + \lambda_z}. \end{aligned} \tag{3.8}$$

Assim, a probabilidade de o servidor A ficar livre antes dos demais é

$$\frac{\lambda_x}{\lambda_x + \lambda_y + \lambda_z}.$$

3.2 A fórmula de Little

A fórmula de Little é utilizada para calcular as medidas de desempenho em sistemas de filas e é utilizada nas seções 3.5.1 e 4.2. Com ela é possível relacionar as medidas L (número médio de usuários no sistema), L_q (número médio de usuários em fila), W (tempo médio de permanência no sistema) e W_q (tempo médio de permanência na fila), sob a hipótese de que o sistema está em equilíbrio. Uma maneira de ilustrar esta fórmula é tomar um exemplo simples, considerando a chegada de três usuários no sistema. A Figura 3.1 representa a chegada de três usuários no sistema até o início do serviço. Assim, o comprimento horizontal de cada barra representa a espera (w_i) dos usuários, enquanto que o comprimento vertical representa o tamanho da fila (q_i) em um dado momento no sistema.

Usuário											Esperas (w_i)
3											5
2											6
1											6
Tempo (t)	1	2	3	4	5	6	7	8	9	10	
Filas (q_i)	1	1	2	2	2	3	2	2	1	1	

Figura 3.1 – Atendimento de três usuários no sistema.

A espera total $\left(\sum_{i=1}^n w_i\right)$ e a fila total $\left(\sum_{i=1}^n q_i\right)$ são duas formas diferentes de

obter a área, no espaço bidimensional usuário x tempo da Figura 3.1. A área é sempre a mesma independente da quantidade de usuários no sistema. Basta observar que o efeito da chegada do terceiro usuário no sistema é aumentar o tempo total de espera em 5 unidades ou aumentar tamanhos da fila em uma unidade em 5 intervalos de tempo. Além disso, a área pode ser calculada a cada vez que um usuário chega no sistema, independente de as observações serem feitas em intervalos discretos e constantes de tempo, bastando multiplicar o intervalo de tempo da espera pelo tamanho da fila.

O tamanho médio da fila e o tempo médio de espera no sistema podem ser

calculados da seguinte forma: $\bar{L}_q = \frac{\sum_{i=1}^n q_i}{t}$ e $\bar{W}_q = \frac{\sum_{i=1}^n w_i}{n}$, respectivamente. Com essa

análise temos: $n \cdot \bar{W}_q = t \cdot \bar{L}_q$, que equivale a $\bar{L}_q = \frac{n}{t} \cdot \bar{W}_q$, em que $\frac{n}{t}$ representa o número médio de chegadas por unidade de tempo. Uma análise equivalente pode ser feita para determinar L e W .

Para um sistema em equilíbrio, depois de n chegadas (n grande) e um tempo suficientemente longo de operação ($t \rightarrow \infty$), as quantidades \bar{W} , \bar{L} e $\frac{n}{t}$ convergem para seu valor médio de forma que: $L = \lim_{t \rightarrow \infty} \bar{L}$, número médio de usuários no sistema; $W = \lim_{t \rightarrow \infty} \bar{W}$, tempo médio de espera no sistema; e $\bar{\lambda} = \lim_{t \rightarrow \infty} \frac{n}{t}$, taxa de entrada de usuários. Assim, sob a hipótese de equilíbrio, temos que:

$$L = \bar{\lambda} W . \quad (3.9)$$

Dados $\bar{\lambda}$, μ e qualquer uma das medidas L , L_s , L_q , W e W_q , pode-se obter as demais a partir da fórmula de Little (Equação (3.9)) e as equações (3.10), (3.11) e (3.12), a seguir:

$$L_q = \bar{\lambda} W_q , \quad (3.10)$$

$$L = L_s + L_q = \frac{\bar{\lambda}}{\mu} + L_q , \quad (3.11)$$

$$W = S + W_q = \frac{1}{\mu} + W_q . \quad (3.12)$$

em que:

$$L_s = \frac{\bar{\lambda}}{\mu} \text{ é o número médio de usuários em serviço.}$$

$$S = \frac{1}{\mu} \text{ é o tempo médio de serviço do sistema.}$$

3.3 A fila $M/M/m$

A fila $M/M/m$, notação de Kendall (KENDALL, 1953), significa que a primeira e a segunda posições indicam as distribuições dos intervalos de tempo entre as chegadas e os tempos de serviço (M indica um processo Markoviano), respectivamente, e a terceira posição é a quantidade de servidores no sistema. As chegadas e os tempos

de serviço possuem médias $E(X) = \lambda$ e $E(S) = \frac{1}{\mu}$, respectivamente. A terceira posição (m) indica o número de servidores idênticos (mesma taxa de serviço). Os usuários esperam em uma única fila com capacidade ilimitada, sendo atendidos pela disciplina FCFS (*First Come First Served*) e o fator de utilização $\rho = \frac{\lambda}{m\mu}$ corresponde à utilização média do sistema ($\rho < 1$). Esse modelo considera que a taxa de entrada de usuários não varia com o estado do sistema, enquanto que as taxas de serviço mudam, dependendo do estado em que o sistema se encontra, da seguinte forma:

$$\lambda_n = \lambda, \text{ para } n = 0, 1, 2, \dots$$

$$\mu_n = \begin{cases} n\mu, & \text{para } n = 1, 2, \dots, m-1 \\ m\mu, & \text{para } n = m, m+1, \dots \end{cases}$$

Para simplificar a análise das Figuras 3.2 e 3.3 não será considerada a chegada de mais de um usuário em um pequeno intervalo de tempo, que acontece com probabilidade $o(\Delta t)$ (como definido na Seção 3.1).

A Figura 3.2 ilustra as possíveis transições de estado do modelo $M/M/m$, quando o número de usuários é menor que o número de servidores ($n < m$). Nestes estados, qualquer usuário que chega no sistema é prontamente atendido. O estado $n + 1$ (no instante t) muda para o estado n com um término de serviço durante Δt , com probabilidade $n\mu\Delta t$. De maneira semelhante, o estado $n - 1$ (no instante t) muda para o estado n com a chegada de uma chamada durante Δt , com probabilidade $\lambda\Delta t$ (esta transição não existe se $n = 0$). Ainda, se no instante t o sistema está no estado n e não ocorre uma chegada ou um término de serviço durante Δt , o sistema permanece no estado n , com probabilidade $1 - \lambda\Delta t - n\mu\Delta t$.

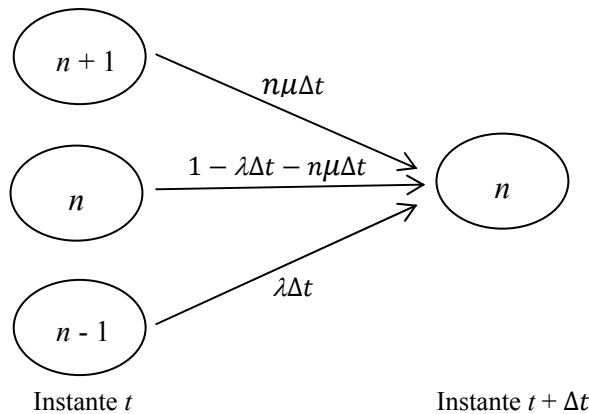


Figura 3.2-Transições de estado do modelo $M/M/m$ a partir do estado n , quando $n < m$.

Quando $n \geq m$ (Figura 3.3), todos os m servidores estão ocupados e se um usuário chega no sistema, ele espera em uma fila simples, operando com disciplina FCFS, como mencionado no início desta Seção. Assim, o estado $n + 1$ (no instante t) muda para o estado n com um término de serviço durante Δt , com probabilidade $m\mu\Delta t$. Caso, no instante t , o sistema esteja no estado n e não ocorra uma chegada ou um término de serviço durante Δt , o sistema permanece no estado n , com probabilidade $1 - \lambda\Delta t - m\mu\Delta t$. O estado $n - 1$ (no instante t) muda para o estado n com a chegada de uma chamada no sistema com probabilidade $\lambda\Delta t$ (da mesma maneira quando $n < m$).

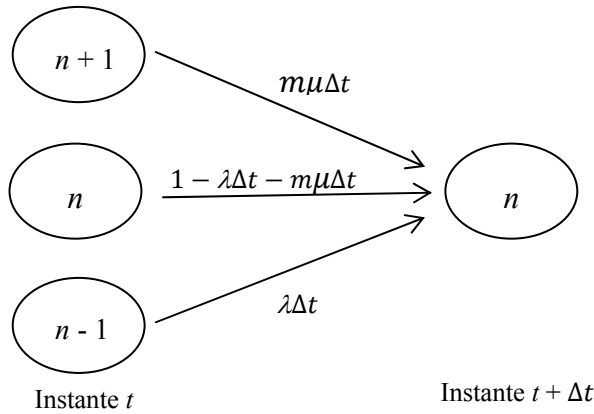


Figura 3.3 - Transições de estado do modelo $M/M/m$ a partir do estado n , quando $n \geq m$.

Dessa forma, a probabilidade de o sistema estar no estado n no instante $t + \Delta t$ é definida por:

$$P_0(t + \Delta t) = P_1(t)\mu\Delta t + P_0(t)[1 - \lambda\Delta t] + o(t), \quad n = 0$$

$$P_n(t + \Delta t) = P_{n+1}(t)(n+1)\mu\Delta t + P_n(t)[1 - (\lambda + n\mu)\Delta t] + P_{n-1}(t)\lambda\Delta t + o(t), \quad n = 1, 2, \dots, m-1$$

$$P_n(t + \Delta t) = P_{n+1}(t)m\mu\Delta t + P_n(t)[1 - (\lambda + m\mu)\Delta t] + P_{n-1}(t)\lambda\Delta t + o(t), \quad n = m, m+1, \dots$$

As equações anteriores podem ser reescritas da seguinte forma:

$$\frac{P_0(t + \Delta t) - P_0(t)}{\Delta t} = \mu P_1(t) - \lambda P_0(t) + \frac{o(t)}{\Delta t}, \quad n = 0$$

$$\frac{P_n(t + \Delta t) - P_n(t)}{\Delta t} = -(\lambda + n\mu)P_n(t) + (n+1)\mu P_{n+1}(t) + \lambda P_{n-1}(t) + \frac{o(t)}{\Delta t}, \quad n = 1, 2, \dots, m-1$$

$$\frac{P_n(t + \Delta t) - P_n(t)}{\Delta t} = -(\lambda + m\mu)P_n(t) + m\mu P_{n+1}(t) + \lambda P_{n-1}(t) + \frac{o(t)}{\Delta t}, \quad n = m, m+1, \dots$$

Fazendo $\Delta t \rightarrow 0$, temos que:

$$\frac{dP_0(t)}{dt} = \mu P_1(t) - \lambda P_0(t), \quad n = 0 \quad (3.13)$$

$$\frac{dP_n(t)}{dt} = -(\lambda + \mu)P_n(t) + (n+1)\mu P_{n+1}(t) + \lambda P_{n-1}(t), \quad n = 1, \dots, m-1$$

$$\frac{dP_n(t)}{dt} = -(\lambda + m\mu)P_n(t) + m\mu P_{n+1}(t) + \lambda P_{n-1}(t), \quad n = m, m+1, \dots$$

As equações diferenciais (3.13) descrevem o sistema ao longo do tempo, permitindo a análise do comportamento transiente do sistema de filas.

A análise em equilíbrio desconsidera o período transiente que ocorre durante o período inicial de operação do sistema. Dessa forma, P_n é a probabilidade de o sistema (em equilíbrio) estar no estado n independentemente de t ($\lim_{t \rightarrow \infty} P_n(t) = P_n$). Estas probabilidades podem ser obtidas fazendo $\frac{dP_n(t)}{dt} = 0$. Assim, o sistema de infinitas equações diferenciais se reduz ao seguinte sistema de equações lineares nas variáveis P_n :

$$\lambda P_0 = \mu P_1 \quad n = 0 \quad (3.14)$$

$$(\lambda + \mu)P_1 = \lambda P_0 + 2\mu P_2 \quad n = 1$$

$$(\lambda + 2\mu)P_2 = \lambda P_1 + 3\mu P_3 \quad n = 2$$

\vdots

$$(\lambda + n\mu)P_n = \lambda P_{n-1} + (n+1)\mu P_{n+1} \quad n = 3, \dots, m$$

$$(\lambda + m\mu)P_n = \lambda P_{n-1} + m\mu P_{n+1} \quad n = m, m+1, \dots$$

Pode-se observar no sistema de equações (3.14) que, em média, a taxa com que o sistema sai do estado n (lado esquerdo das equações) é igual a taxa com que o sistema entra no estado n (lado direito das equações). Cada equação do Sistema (3.14) pode ser escrita de forma recursiva da seguinte forma:

$$P_1 = \frac{\lambda}{\mu} P_0, \quad (3.15)$$

$$P_2 = \frac{\lambda P_1 + \mu P_1 - \lambda \frac{\mu}{\lambda} P_1}{2\mu} = \frac{\lambda}{2\mu} P_1, \text{ e assim por diante.}$$

A relação de recursão observada no sistema (3.14) é válida apenas para fila simples, em que os estados se comunicam 2 a 2, e mostra que o estado $n = 0$ entra em equilíbrio com o estado $n = 1$; o estado $n = 1$ entra em equilíbrio com o estado $n = 2$; e assim por diante. A Figura 3.4 mostra a transição de estados em um sistema $M/M/m$ em equilíbrio, para $n = 0, 1, 2, \dots$.

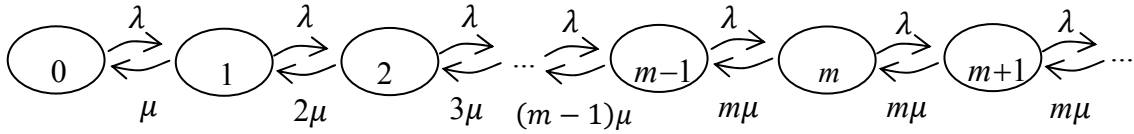


Figura 3.4- Transição de estados do modelo $M/M/m$.

O sistema (3.15) pode ser resolvido computacionalmente, definindo a função auxiliar $F(n) = \frac{P_n}{P_0}$ (Equação (3.16)). A relação de recorrência existente nas equações do sistema (3.15) é mantida na Equação (3.16) de $F(n)$ [Pode-se notar que $F(0) = 1$].

$$F(n) = \begin{cases} 1 & \text{se } n = 0 \\ \frac{\lambda}{n\mu} F(n-1) & \text{se } n \leq m \\ \frac{\lambda}{m\mu} F(n-1) & \text{se } n > m \end{cases} \quad (3.16)$$

Temos que $P_0 \sum_{n=0}^{\infty} F(n) = 1$ e P_0 e P_n podem ser encontrados da seguinte

forma:

$$P_n = \frac{F(n)}{\sum_{n=0}^{\infty} F(n)}.$$

O número médio de usuários na fila (L_q), Equação (3.17), é obtido a partir das probabilidades dos estados P_n . As demais medidas de desempenho (L , L_s , L_q , W e W_q) podem ser obtidas utilizando a fórmula de Little (3.9) e as equações (3.10), (3.11) e (3.12).

$$L_q = \sum_{n=m}^{\infty} (n-m) P_n. \quad (3.17)$$

3.4 A fila $M/M/m/C$

O modelo $M/M/m/C$ ($m \leq C$) difere do modelo $M/M/m$ apenas pela limitação c do número de usuários presentes no sistema (em fila e em serviço), resultando em tamanho máximo para a fila ($C - m$). O diagrama de transição de estados que representa o sistema em equilíbrio pode ser visto na Figura 3.5. As taxas de transição que representam este sistema são:

$$\lambda_n = \begin{cases} \lambda, & \text{para } n = 0, 1, 2, \dots, m-1 \\ 0, & \text{para } n = m, m+1, \dots, C \end{cases}$$

$$\mu_n = \begin{cases} n\mu, & \text{para } n = 0, 1, 2, \dots, m-1 \\ m\mu, & \text{para } n = m, m+1, \dots, C \end{cases}$$

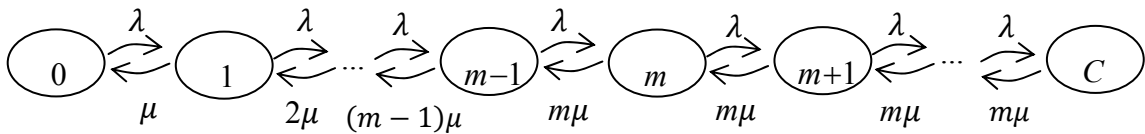


Figura 3.5 - Transição de estados do modelo $M/M/m/c$.

Utilizando o resultado da seção 3.3, segundo a qual, em média, a taxa com que o sistema sai do estado n é igual a taxa com que o sistema entra no estado n , a distribuição de equilíbrio do modelo é dada por:

$$\begin{aligned} \lambda P_0 &= \mu P_1 & n=0 \\ (\lambda + \mu) P_1 &= \lambda P_0 + 2\mu P_2 & n=1 \\ (\lambda + 2\mu) P_2 &= \lambda P_1 + 3\mu P_3 & n=2 \end{aligned} \quad (3.18)$$

$$\vdots$$

$$(\lambda + n\mu)P_n = \lambda P_{n-1} + (n+1)\mu P_{n+1} \quad n = 3, 4, \dots, m$$

$$(\lambda + m\mu)P_n = \lambda P_{n-1} + m\mu P_{n+1} \quad n = m, m+1, \dots, C$$

O sistema (3.18) pode ser resolvido computacionalmente utilizando a função auxiliar (3.16) impondo a condição $n \leq C$ (fila finita). Para esse modelo, a probabilidade de perda é uma medida de desempenho importante a ser considerada, e representa a probabilidade de o sistema estar com c usuários ($P_{\text{perda}} = P_C$). A taxa de entrada de usuários ($\bar{\lambda} < \lambda$) utilizada na fórmula de Little é definida por:

$$\bar{\lambda} = \lambda \sum_{n=0}^{c-1} P_n + 0P_C = \lambda(1 - P_C).$$

Nesse modelo, o fator de utilização ρ não corresponde à utilização média do sistema, a qual é dada por $\frac{L_s}{m} = \frac{\bar{\lambda}}{m\mu}$. O número médio de usuários na fila é dado pela Equação (3.19). Pode-se notar que, se $c \rightarrow \infty$ e $\rho < 1$, as expressões de P_0, P_n e L_q se reduzem às do modelo $M/M/m$, discutido na Seção 3.3. As demais medidas L, L_s, L_q, W e W_q podem ser obtidas a partir da fórmula de Little (3.9) e das equações (3.10), (3.11) e (3.12).

$$L_q = \sum_{n=m}^C (n-m)P_n \quad (3.19)$$

3.5 O modelo hipercubo clássico

Conforme mencionado no Capítulo 2, o modelo hipercubo é um modelo descritivo utilizado como uma ferramenta para análise e planejamento de sistemas de emergência urbanos. Além de considerar incertezas quanto à origem dos chamados, tempos de serviço e disponibilidade dos servidores, o modelo aborda complexidades geográficas e temporais da região, com base em filas espacialmente distribuídas. Ele pode analisar tanto sistemas coordenados como centralizados, quando o usuário liga para uma central solicitando algum tipo de serviço e um servidor se desloca até o cliente.

Basicamente, a idéia é expandir o espaço de estados de um sistema de fila $M/M/m$ a fim de representar cada servidor individualmente, podendo incluir políticas de despacho mais complicadas. A solução do modelo é dada partindo-se da construção do conjunto de equações de equilíbrio para o sistema. Os resultados baseiam-se nos valores das probabilidades de estado do sistema, possibilitando o cálculo de medidas de desempenho, tais como: cargas de trabalho dos servidores, tempo médio de resposta do sistema ou de cada servidor, frequência de atendimento de cada servidor em cada região, entre outras. Algumas destas hipóteses podem ser alteradas, como, por exemplo, múltiplo despacho e *backup* parcial, como em Chelst e Barlach (1981), Mendonça e Morabito (2000), Iannoni (2005) e Iannoni *et. al.* (2008a, 2008b).

O modelo hipercubo baseia-se na divisão da região atendida pelo sistema em átomos geográficos (regiões de demanda). Cada átomo é considerado uma fonte de chamados pontual e independente das demais e o atendimento a cada átomo é realizado por servidores que estão distribuídos na região. A localização dos servidores deve ser conhecida ou estimada por meio de probabilidade geométrica. Se um servidor estiver ocupado, outros servidores poderão atender ao chamado, mesmo que seja fora de sua área preferencial, prevalecendo a cooperação entre os servidores.

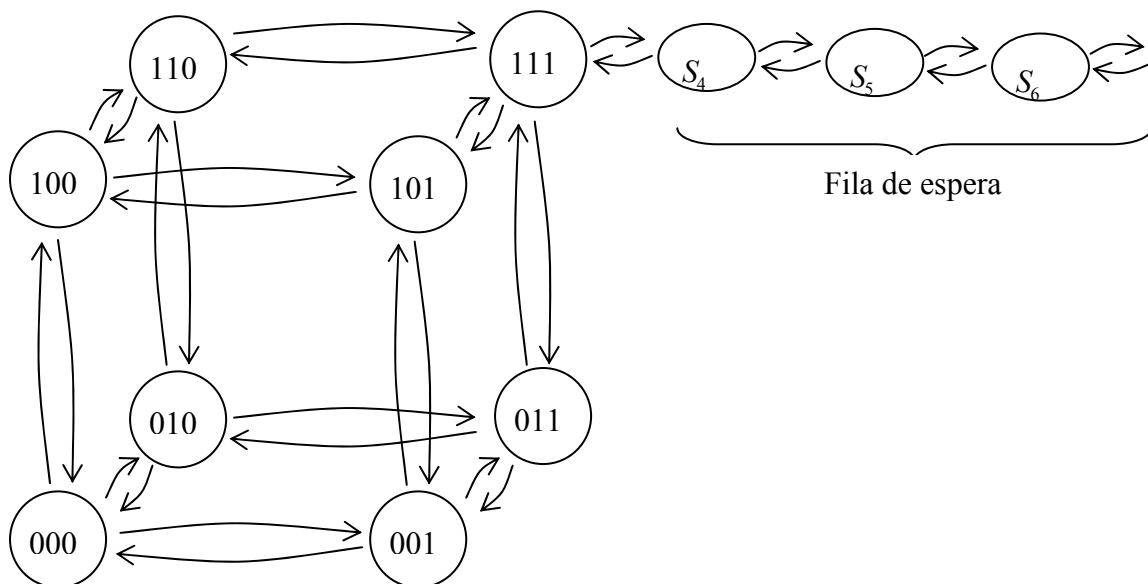


Figura 3.6 – Estados do sistema com três servidores.

A disponibilidade dos servidores é representada por meio do espaço de estados dos servidores. Um estado particular do sistema sem fila, é dado pela lista dos

servidores que estão livres ou ocupados. Sejam $\{000\}$, $\{001\}$, $\{010\}$, ..., $\{111\}$ os $2^3 = 8$ possíveis estados do sistema, em que os 0's e 1's indicam se os servidores estão livres ou ocupados, respectivamente. Por exemplo, o estado $\{011\}$ representa o estado em que o servidor 1 está livre e os servidores 2 e 3 estão ocupados ($\{011\}$ descreve o estado dos servidores da esquerda para a direita). Assim, o espaço de estados de um sistema com três servidores pode ser representado por um cubo; no caso de haverem mais de três servidores, temos um hipercubo. A Figura 3.6 mostra o espaço de estados para sistemas com $m = 3$ servidores.

O modelo trata tanto de sistemas em que não é permitida a formação de fila, como aqueles em que, quando todos os servidores estão ocupados, os chamados que chegam esperam em uma fila, por meio da qual os usuários são atendidos à medida em que os servidores tornam-se livres segundo a disciplina FCFS. O modelo hipercubo estabelece uma relação com o modelo clássico $M/M/m$ (discutido na Seção 3.3), caso os servidores sejam idênticos, com a mesma taxa de atendimento ($\mu_1 = \mu_2 = \dots = \mu_m$) (LARSON e ODoni, 2007). Isso permite relacionar as probabilidades de estado do modelo hipercubo com as probabilidades de estado do modelo $M/M/m$, como a seguir:

$$P_{\{00\dots 0\}} = P_{M/M/m} \{S_0\} \quad (3.20)$$

$$\begin{aligned} P_{\{10\dots 0\}} + P_{\{01\dots 0\}} + \dots + P_{\{00\dots 1\}} &= P_{M/M/m} \{S_1\} \\ &\vdots \\ P_{\{11\dots 1\}} &= P_{M/M/m} \{S_n\} \end{aligned}$$

em que $P_{M/M/m} \{S_n\}$ representa a probabilidade de um sistema $M/M/m$ encontrar-se no estado $n = 0, 1, \dots$. Essas equações são denominadas equações dos hiperplanos definidos pelo modelo. Além de o modelo hipercubo considerar servidores heterogêneos, ele se diferencia, ainda, do modelo $M/M/m$ por ser capaz de considerar políticas de despacho particulares para cada servidor.

Conforme mencionado no Capítulo 2, o modelo hipercubo é um modelo descritivo que não permite, se aplicado isoladamente, a solução direta de problemas de localização a fim de, por exemplo, diminuir o tempo médio de resposta ao usuário ou diminuir a carga de trabalho das ambulâncias. Porém, é um modelo que permite estimar

medidas de desempenho importantes, possibilitando a análise de cenários alternativos.

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

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

Conforme descrito no Capítulo 2, algumas dessas hipóteses podem ser desconsideradas. Um exemplo disto é o modelo proposto neste trabalho, descrito no Capítulo 4.

A seguir, é apresentado o modelo hipercubo por meio de um exemplo simples, resolvido em Chiyoshi *et. al* (2001). Considere um sistema de emergência

operando em uma região representada por três átomos, utilizando política de despacho de preferência fixa, mostrada na Tabela 3.1.

Átomo	Matriz de Despachos		
	Preferências		
	1º	2º	3º
1	1	2	3
2	2	3	1
3	3	1	2

Tabela 3.1– Matriz de Preferências de despacho.

A solução do modelo é dada pela construção das equações de equilíbrio do sistema, que são definidas supondo-se que o sistema atinja o equilíbrio. Para cada estado do sistema, o fluxo que entra neste estado deve ser igual ao fluxo que sai dele. Em um sistema não saturado, com capacidade de fila infinita, as probabilidades de estado do modelo hipercubo são calculadas a partir das equações de balanço, construídas a partir dos oito possíveis estados, descritos anteriormente nesta mesma seção.

Quando o sistema está no estado $\{000\}$ (sistema vazio), ele passa para o estado $\{100\}$ quando ocorre um chamado com origem no átomo 1, com taxa de ocorrência λ_1 . O mesmo acontece com o estado $\{010\}$, com taxa λ_2 , e para o estado $\{001\}$, com taxa λ_3 . Dessa forma, a taxa total de transição do estado $\{000\}$ para outros estados é $\lambda = \lambda_1 + \lambda_2 + \lambda_3$.

No sentido contrário, o estado $\{000\}$ pode ser alcançado a partir do estado $\{100\}$ quando o servidor 1 termina o atendimento, com taxa μ_1 ; da mesma forma, a partir do estado $\{010\}$, com taxa μ_2 ; e a partir de $\{001\}$, com taxa μ_3 . Podemos obter a equação de equilíbrio para o estado $\{000\}$ a partir da definição obtida na Seção 3.3, de que “a taxa com que o sistema entra no estado n deve ser igual a taxa com que o sistema sai do estado n ”, da seguinte forma:

$$\lambda P_{\{000\}} = \mu_1 P_{\{100\}} + \mu_2 P_{\{010\}} + \mu_3 P_{\{001\}} \quad (3.21)$$

Com esse procedimento, podemos obter as equações para os estados seguintes, obtendo o conjunto de equações (3.22).

$$\{000\} \rightarrow \lambda P_{\{000\}} = \mu_1 P_{\{100\}} + \mu_2 P_{\{010\}} + \mu_3 P_{\{001\}} \quad (3.22)$$

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

Em que:

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

Pela condição de equilíbrio do sistema, a transição entre os estados $\{111\}$ e $\{S_4\}$ devem ser iguais, de forma que $\lambda P_{111} = \mu P_{S_4}$. Caso essa condição não aconteça, o sistema está na fase transiente e a cauda estaria em crescimento. Assim, a oitava equação do Sistema (3.22) pode ser escrita na forma:

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

Chiyoshi *et al* (2000) mencionam que o sistema (3.22) escrito na forma matricial $Ax = 0$ é um sistema linear homogêneo indeterminado. Isso ocorre porque as equações apenas impõem condições de equilíbrio para cada estado do sistema $\{000\}, \{001\}, \{010\}, \{100\}, \{011\}, \{101\}, \{110\}, \{111\}$, mas nada especifica sobre como a massa total de probabilidade se distribui entre estes estados e os estados da cauda. Uma maneira de tornar o sistema determinado é a substituição de uma das equações do sistema por uma equação de normalização, considerando que $\sum_{n=0}^N P_n = 1$, sendo que N é o

número de estados possíveis para o sistema. A equação de normalização é dada por:

$$P_{\{000\}} + P_{\{001\}} + P_{\{010\}} + P_{\{100\}} + \dots + P_{\{111\}} + P_{\{S_4\}} + P_{\{S_5\}} + \dots = 1 \quad (3.23)$$

A partir do estado $\{111\}$ temos uma fila infinita onde os estados se comunicam 2 a 2, assim como o modelo $M/M/m$ descrito na Seção 3.3. O estado $\{111\}$ está em equilíbrio à esquerda com o hipercubo e para ficar em equilíbrio com o S_4 , tem-se que $(\lambda)P_{\{111\}} = \mu P_{\{S_4\}}$. É necessário que $P_{\{S_4\}} = \left(\frac{\lambda}{\mu}\right)^2 P_{\{111\}}$ para que o estado S_4 fique

em equilíbrio com o estado S_5 , e assim por diante. Dado que $\frac{\lambda}{\mu} = \rho$, em que $\rho < 1$,

temos as equações simplificadas para:

$$\triangleright P_{\{S_4\}} = \rho P_{\{111\}}.$$

De forma similar, para a transição dos estados correspondentes a dois e três usuários na fila S_5 e S_6 , tem-se:

$$\triangleright P_{\{S_5\}} = \rho^2 P_{\{111\}};$$

$$\triangleright P_{\{S_6\}} = \rho^3 P_{\{111\}};$$

$$\vdots$$

$$\triangleright P_{\{S_n\}} = \rho^n P_{\{111\}};$$

$$\vdots$$

Assim, somando as probabilidades dos estados nas quais todos os servidores estão ocupados, tem-se:

$$P_{\{111\}} + P_{\{S_4\}} + \dots + P_{\{S_n\}} + \dots = P_{\{111\}} + \rho P_{\{111\}} + \rho^2 P_{\{111\}} + \dots + \rho^n P_{\{111\}} + \dots = P_{\{111\}} \sum_{j=0}^{\infty} \rho^j. \quad (3.24)$$

Pode-se notar que, como $\rho < 1$, $\sum_{j=0}^{\infty} \rho^j = \frac{1}{(1-\rho)}$ (uma série geométrica de

razão ρ), tem-se:

$$P_{\{111\}} + P_{\{s_4\}} + \dots + P_{\{s_n\}} + \dots = \frac{P_{\{111\}}}{(1-\rho)}. \quad (3.25)$$

Substituindo na equação de normalização, tem-se:

$$P_{\{000\}} + P_{\{100\}} + P_{\{010\}} + P_{\{001\}} + \dots + P_{\{011\}} + \frac{P_{\{111\}}}{(1-\rho)} = 1$$

Dessa forma, os estados do modelo hipercubo podem ser calculados separadamente da cauda. Diversas medidas de desempenho são calculadas a partir das probabilidades de estado, obtidas pela solução das equações de equilíbrio do sistema. Portanto, essas medidas auxiliam a análise do sistema sob a hipótese de que o sistema está em equilíbrio.

➤ Carga de trabalho (*workload*) é a fração de tempo em que o servidor está ocupado e é calculada somando-se as probabilidades dos estados em que este servidor estiver ocupado.

$$\rho_i = \sum_{\{B:b_i=1\}} P_B + P_Q, \text{ em que:} \quad (3.26)$$

- ✓ ρ_i é a carga de trabalho (*workload*) do servidor i ($i = 1, 2, \dots, m$);
- ✓ $\sum_{\{B:b_i=1\}} P_B$ é a soma das probabilidades dos estados (de $\{000\}$ a $\{111\}$), em que o servidor i está ocupado ($b_i = 1$);
- ✓ P_Q é a probabilidade de fila ($P_Q = P_{\{s_4\}} + P_{\{s_5\}} + \dots$).

Para o caso particular do exemplo de um sistema com três servidores, temos que:

$$\rho_1 = P_{\{001\}} + P_{\{011\}} + P_{\{101\}} + P_{\{111\}} + P_Q$$

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

$$\rho_3 = P_{\{100\}} + P_{\{101\}} + P_{\{110\}} + P_{\{111\}} + P_Q$$

➤ A frequência de despacho indica a fração de despachos no sistema que é atendida pelo servidor i ($i = 1, 2, \dots, m$) no átomo j ($j = 1, 2, \dots, N_A$), e é dada pela soma

de duas parcelas: $f_{ij}^{[nq]}$, fração de despachos em que o servidor i é enviado ao átomo j , mas não implica tempo de espera em fila; $f_{ij}^{[q]}$, fração de despachos em que o servidor i é enviado ao átomo j , e implica tempo de espera em fila.

$$f_{ij} = f_{ij}^{[nq]} + f_{ij}^{[q]} = \frac{\lambda_j}{\lambda} \sum_{B \in E_{ij}} P_B + \frac{\lambda_j}{\lambda} P_Q \cdot \frac{\mu_i}{\mu}. \text{ Em que,} \quad (3.27)$$

A primeira parcela da Equação (3.27) $\left(f_{ij}^{[nq]}\right)$ é dada pela probabilidade de surgir um chamado no átomo j antes dos demais (dado que os chamados chegam a partir de um processo de Poisson, Seção 3.1); $\frac{\lambda_j}{\lambda}$, e a probabilidade de o servidor i ser o primeiro servidor disponível na lista de preferência de despacho do átomo j ; $\sum_{B \in E_{ij}} P_B$ (em que E_{ij} é o conjunto dos estados nos quais o servidor i é o primeiro servidor disponível na lista de despacho do átomo j).

A segunda parcela da Equação (3.27) $\left(f_{ij}^{[q]}\right)$ é dada pela probabilidade de surgir um chamado no átomo j antes dos demais, $\frac{\lambda_j}{\lambda}$. Como todos os servidores estão ocupados, a probabilidade de o servidor i ser o primeiro a ficar livre (todos os servidores possuem tempos de serviço independentes e exponencialmente distribuídos) é dada por $\frac{\mu_i}{\mu}$. Deve-se ter ainda o sistema em estado de saturação, com probabilidade P_Q . $\left(P_Q = P_{\{111\}} + P_Q\right)$.

➤ O tempo médio de viagem no sistema é:

$$\bar{T} = \sum_{i=1}^m \sum_{j=1}^{N_A} f_{ij}^{[nq]} t_{ij} + P_Q \bar{T}_Q. \quad (3.28)$$

em que,

➤ t_{ij} , é o tempo médio de viagem do servidor i ao átomo j ;

- $\bar{T}_Q = \sum_{p=1}^{N_A} \sum_{j=1}^{N_A} \frac{\lambda_p \lambda_j}{\lambda^2} \tau_{pj}$, é tempo médio de viagem para chamados em fila;
- $\frac{\lambda_p}{\lambda}$ e $\frac{\lambda_j}{\lambda}$, correspondem, respectivamente, à probabilidade de uma chamada que incorre em algum tempo de espera em fila ser gerada no átomo j , e à probabilidade de esta chamada ser atendida por um servidor localizado no átomo p ;
- Utilizando uma das propriedades da distribuição exponencial (Seção 3.1), na qual, $\frac{\mu_i}{\mu}$ indica a probabilidade de o servidor i ser o primeiro a terminar o serviço, considerando variáveis aleatórias independentes e exponencialmente distribuídas.
- τ_{pj} , é o tempo médio de viagem entre os átomos p e j .

- O tempo médio de viagem ao átomo j é calculado por:

$$\bar{T}_j = \frac{\sum_{i=1}^m f_{ij}^{[nq]} t_{ij}}{\sum_{i=1}^m f_{ij}^{[nq]}} (1 - P_{Q'}) + \sum_{p=1}^{N_A} \frac{\lambda_p}{\lambda} \tau_{pj} P_{Q'}. \quad (3.29)$$

- O tempo médio de viagem do servidor i pode ser aproximado por:

$$\overline{TU}_i = \frac{\sum_{j=1}^{N_A} f_{ij}^{[nq]} t_{ij} + (T_Q P_{Q'} / m)}{\sum_{j=1}^{N_A} f_{ij}^{[nq]} + (P_{Q'} / m)}. \quad (3.30)$$

Outras medidas de desempenho podem ser definidas e calculadas, como a fração de despachos do servidor i como *backup*, a fração de despachos de *backup* para o átomo j e a fração total de despachos *backup* no sistema (Larson e Odoni, 2007).

Resolver o modelo hipercubo significa encontrar a solução de um sistema

linear (um exemplo com três servidores é mostrado no sistema (3.22)) com $O(2^m)$ equações, sendo m o número de servidores, em um sistema com três servidores teremos um sistema com oito equações. Mesmo para pequenos valores de m , a solução do sistema pode ser inviável computacionalmente, sendo necessária a utilização de métodos aproximados. Para resolver o sistema de forma exata pode-se utilizar o método de Gauss-Siedel, que é um processo iterativo que utiliza os valores das variáveis mais recentes para determinar o valor das variáveis atuais.

Quando o método exato é inviável computacionalmente, utiliza-se o método aproximado de Larson (1975), que envolve a resolução de um sistema não linear de m equações que tem como incógnitas as m taxas de ocupação dos servidores, e não as 2^m probabilidades de estado do método exato. Nesse caso, as frequências de despacho devem ser calculadas a partir das taxas de ocupação dos servidores. O método aproximado de Larson admite que os servidores são homogêneos, isto é, todos têm a mesma taxa de serviço. Outra característica do método aproximado é que ele utiliza a derivação de um fator Q baseando-se em um processo de amostragem aleatória dos servidores em um sistema $M/M/m$.

Quando as taxas de ocupação dos servidores não são muito diferentes e muitos vetores de preferência de despachos simulam, no conjunto, uma seleção aleatória de servidores, o método aproximado de Larson é utilizado como uma aproximação do modelo hipercubo com servidores diferentes.

3.6 Filas com classes de usuários e distribuições exponenciais

Todos os modelos descritos nas Seções 3.3 a 3.5 foram determinados pela utilização da disciplina FCFS. Esses modelos não diferenciam os usuários de acordo com o tipo ou duração do serviço. Na prática, muitos sistemas diferenciam usuários, de forma que eles sejam separados em classes com diferentes prioridades. Por exemplo, em plantões de hospitais ou postos de saúde, pacientes que correm risco de vida devem ser atendidos com maior prioridade do que outros pacientes menos graves (que não correm risco de vida).

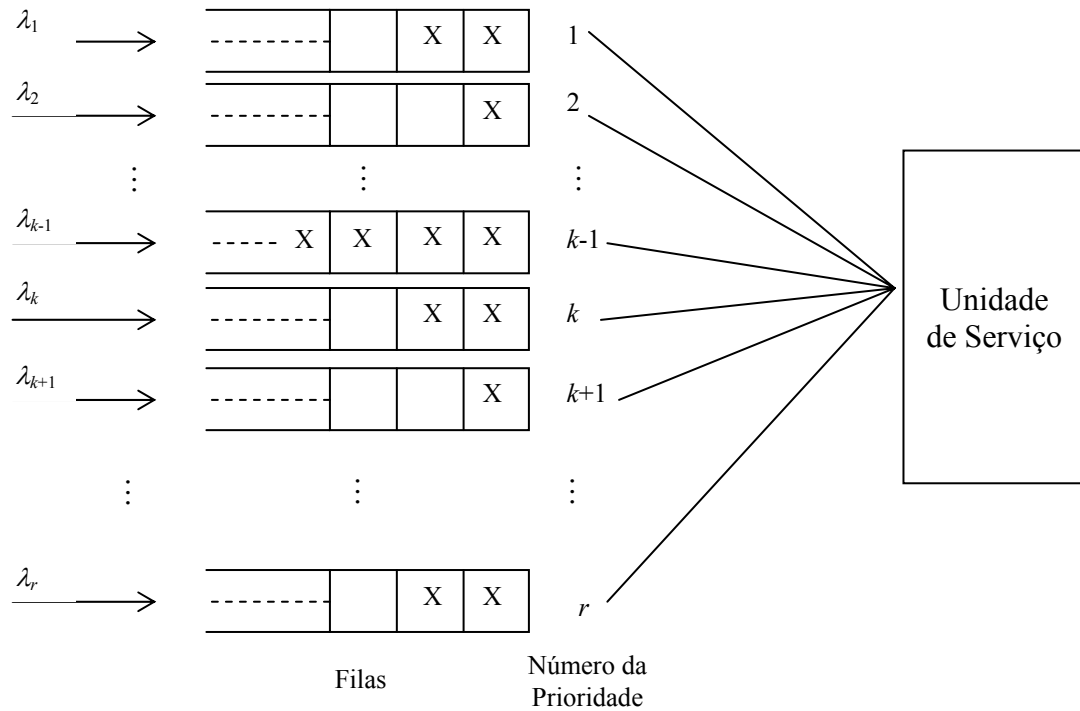


Figura 3.7 – Representação de um sistema de filas com r classes e disciplina de prioridade.

Em um sistema de filas com r classes de usuários, como o da Figura 3.7, cada classe é denominada por um número k , $k = 1, 2, \dots, r$, em que o menor número representa a classe com prioridade mais alta, enquanto o maior representa a classe com prioridade mais baixa. Pode-se assumir que cada uma das r filas anda com base em uma disciplina FCFS. Uma característica importante de um sistema utilizando disciplina de prioridade é que um usuário de uma classe com prioridade menor nunca pode ser servido se um usuário de uma classe com prioridade maior estiver presente no sistema.

3.6.1 Prioridades sem interrupção em um sistema $M_r/M/m$

Os modelos que consideram disciplinas de filas com prioridades são divididos em dois tipos: prioridades com e sem interrupção. Em ambos os casos, o sistema sempre escolhe o próximo usuário na fila que representa a classe de prioridade mais alta. A diferença é que, em sistemas utilizando prioridade com interrupção, um usuário com prioridade mais alta nunca pode esperar em favor de um usuário com prioridade mais baixa, mesmo estando em serviço. Nesse caso, usuários que saem da unidade de serviço pela chegada de um usuário de classe com prioridade mais alta,

podem voltar a serem servidos a partir do ponto que o serviço foi interrompido ou recomeçar do início. Nota-se que, para usuários com tempo de serviço distribuídos exponencialmente, estes dois casos são indistinguíveis, em consequência da propriedade da falta de memória da distribuição exponencial.

Nos modelos que consideram prioridades sem interrupção, o próximo usuário a ser atendido deve esperar o usuário em atendimento terminar o serviço, independente da sua classe. O modelo sem interrupção é bastante utilizado em SAE's, quando uma ambulância que está em serviço deve terminar o atendimento de um usuário antes de começar outro.

No sistema $M_r/M/m$, a chegada de usuários da classe k ocorre segundo o processo de Poisson, com taxa de chegada λ_k . Há m servidores idênticos com taxa de serviço $\frac{1}{\mu}$, que operam em regime de prioridade sem interrupção e capacidade da fila infinita. A taxa de utilização da classe k pode ser escrita como $\rho_k = \frac{\lambda_k}{m\mu}$ e a taxa de utilização do sistema é dada por $\rho = \rho_1 + \rho_2 + \dots + \rho_r$.

Sob as condições descritas nesta seção, pode-se derivar uma expressão para calcular o tempo médio de espera na fila para um usuário em uma classe com prioridade k (W_{qk}). Considerando a chegada de um usuário originário da classe k do sistema, pode-se obter:

$$W_{qk} = W_0 + \frac{1}{m\mu} \sum_{i=1}^k L_{qi} + \frac{1}{m\mu} \sum_{i=1}^{k-1} M_i \quad (3.31)$$

em que:

W_0 é o tempo médio de espera para um dos servidores tornar-se livre quando ocorre a chegada de um novo usuário (da classe k) no sistema, temos que (assumindo $\lambda_1 + \lambda_2 + \dots + \lambda_r < m\mu$):

$$W_0 = \frac{1}{m\mu} \sum_{n=m}^{\infty} P_n. \quad (3.32)$$

$L_{qi} = \lambda_i W_{qi}$ é o número médio de usuários na fila com prioridade maior ou igual k ($i \geq k$) que estão esperando no instante que o usuário da classe k chega no sistema.

$M_i = \lambda_i W_{qk}$ é o número médio de usuários com prioridade maior que k ($i > k$) que chegaram no sistema depois dele. Pode-se notar que W_{qk} ainda é desconhecido.

Substituindo as duas expressões acima na Equação (3.31), temos que:

$$W_{qk} = W_0 + \sum_{i=1}^k \rho_i W_{qi} + W_{qk} \sum_{i=1}^{k-1} \rho_i \quad (3.33)$$

Considerando a Equação (3.33), o primeiro somatório representa o tempo médio de espera de um usuário classe k , uma vez que todos os usuários que chegaram primeiro com prioridade maior ou igual à dele são atendidos prioritariamente. O segundo somatório representa o tempo médio de espera do usuário da classe k presente no sistema, uma vez que podem ocorrer chegadas de usuários com prioridade maior que a dele, que serão atendidos preferencialmente. Resolvendo a Equação (3.33) para W_{qk} , obtemos:

$$W_{qk} = \frac{W_0 + \sum_{i=1}^k \rho_i W_{qi}}{1 - \sum_{i=1}^{k-1} \rho_i}, \text{ para } k = 1, 2, \dots, r. \quad (3.34)$$

A Equação (3.34) pode ser resolvida recursivamente, começando com W_{q1} , W_{q2} , Ou seja, começando com W_{q1} , temos que:

$$W_{q1} = \frac{W_0}{1 - \rho_1}, \text{ para } k = 1.$$

$$W_{q2} = \frac{W_0 + \sum_{i=1}^2 \rho_i W_{qi}}{1 - \rho_1}, \text{ para } k = 2.$$

Escrevendo a equação acima em função de W_{q2} , temos que:

$$W_{q2} - \rho_1 W_{q2} - \rho_2 W_{q2} = W_0 + \rho_1 W_{q1},$$

Substituindo W_{q1} , pode-se obter a expressão de W_{q2} :

$$W_{q2} \left(1 - \sum_{i=1}^2 \rho_i \right) = W_0 \left(1 + \frac{\rho_1}{1 - \rho_1} \right),$$

$$W_{q2} = \frac{W_0}{\left(1 - \sum_{i=1}^2 \rho_i\right)(1 - \rho_1)}.$$

Assim, de forma geral, para uma classe k pode-se escrever:

$$W_{qk} = \frac{W_0}{(1 - a_{k-1})(1 - a_k)}, \text{ para } k = 1, 2, 3, \dots, r. \quad (3.35)$$

$$\text{em que } a_k = \sum_{i=1}^k \rho_i, \quad a_0 = 0 \quad \text{e} \quad W_0 = \frac{1}{m\mu} \sum_{n=m}^{\infty} P_n.$$

A fórmula de Little (Equação (3.9)) é válida para qualquer disciplina de fila, e o mesmo vale para \bar{L}, \bar{L}_q e \bar{W}_q (equações (3.10), (3.11) e (3.12)), como discutido na Seção 3.2. A partir do tempo médio de espera na fila para cada classe k , obtido pela Equação (3.35), pode-se obter expressões para L_{qk}, W_k e L_k . Por exemplo:

$L_{qk} = \lambda_k W_{qk}$ é o número médio de usuários da classe k na fila;

$L_k = L_{qk} + \frac{\bar{\lambda}_k}{m\mu}$ é o número médio de usuários da classe k no sistema; e

$W_k = W_{qk} + \frac{1}{m\mu}$ é o tempo médio de espera da classe k no sistema.

3.6.2 Prioridades sem interrupção no modelo hipercubo

Uma maneira de diferenciar usuários e representar prioridades no modelo hipercubo (discutido na Seção 3.5) é subdividir os átomos em subátomos (“*layering*”). Esta abordagem foi utilizada por Takeda (2000) e Iannoni (2005), sem prejuízo da análise. Como mencionado na Seção 2.1, Takeda (2000) subdividiu cada átomo geográfico em dois subátomos (não geográficos) para representar usuários mais graves (pacientes que correm risco de vida) e usuários menos graves (pacientes que não correm risco de vida).

Essa abordagem também é utilizada nesta tese (Capítulo 4) e possibilita representar prioridades na política de despacho das ambulâncias, de forma que usuários mais graves sejam atendidos preferencialmente por um VSA, enquanto que usuários

menos graves sejam atendidos preferencialmente por um VSB. Caso todos os servidores estejam ocupados, o primeiro servidor a ficar disponível atende o chamado, independente do tipo do chamado. Porém, em sistemas congestionados (como o SAMURP) que permitem a formação de fila de espera, essa abordagem torna-se aproximada, pois o modelo hipercubo clássico considera fila com disciplina de atendimento FCFS.

Assim, o próximo capítulo trata o caso em que, além de utilizar a estratégia de *layering* para representar prioridades no modelo hipercubo, os usuários também devem ser diferenciados na cauda, a fim de tratar sistemas mais congestionados de forma exata.

4 Extensão do modelo hipercubo para análise do SAMU-RP

Este capítulo trata da modelagem de um sistema de filas considerando diferentes classes de usuários e disciplina de prioridade na fila. Esta é uma das particularidades do SAMU-RP, como pode ser observado no exemplo da Figura 4.1, que ilustra uma extensão do modelo hipercubo clássico, de forma que se todos os servidores estiverem ocupados, o atendimento será feito utilizando uma fila de espera com disciplina de prioridade e fila finita. No final do capítulo, é enfatizada a relação do modelo hipercubo proposto com o modelo $M/M/m/c$ e, ainda, é feito um procedimento em pseudocódigo para generalização do modelo proposto.

4.1 Exemplo ilustrativo

A fim de apresentar o modelo considerando a disciplina de fila com prioridade, é feito um exemplo ilustrativo considerando três classes de usuários. O objetivo é representar um sistema simplificado, mas com as mesmas características do sistema descrito na Seção 2.1.2: a , chamados de emergência; b , chamados de urgência moderada; c , chamados de urgência leve.

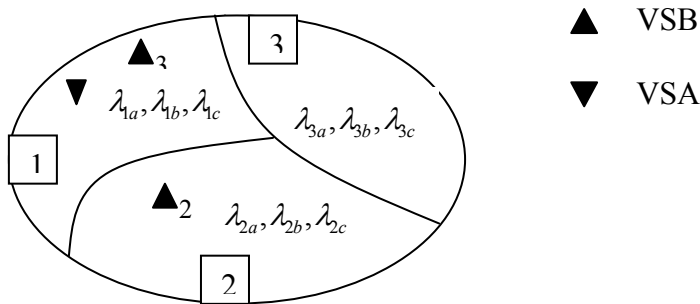


Figura 4.1 – Átomos geográficos

Para a aplicação do modelo hipercubo é necessário considerar algumas premissas, escolhidas de forma a tornar o exemplo ilustrativo o mais parecido possível do SAMU-RP:

- i. A região é particionada em três átomos geográficos ($N_A = 3$), como

mostrado na Figura 4.1 (átomos 1, 2 e 3).

ii. O sistema tem $N = 3$ servidores, localizados da seguinte forma:

- Os servidores 1 (VSA) e 3 (VSB) estão localizados no átomo 1; e,
- O servidor 2 (VSB) está localizado no átomo 2.

iii. $r = 3$ classes de usuários.

iv. Cada átomo geográfico i ($i = 1, 2, \dots, N_A$) foi subdividido em três camadas ou subátomos (não geográficos), que representam as classes de prioridades (por exemplo, o átomo 1 é dividido nos subátomos $1a$, $1b$ e $1c$, representando as classes a , b e c no átomo 1, conforme Figura 4.1). Seja $D = \{a, b, c\}$ o conjunto destas classes de usuários. Pode-se notar que, desta maneira, o sistema, originalmente com $N_A=3$ átomos geográficos, passa a ser analisado como um sistema alternativo com $N_A \times |D| = 3 \times 3 = 9$ subátomos. Admite-se que em cada subátomo, os chamados chegam de acordo com o Processo de Poisson. As taxas dos átomos correspondem a soma das taxas dos respectivos subátomos:

- $\lambda_1 = \lambda_{1a} + \lambda_{1b} + \lambda_{1c}$ é a taxa de chegada de usuários do átomo 1;
- $\lambda_2 = \lambda_{2a} + \lambda_{2b} + \lambda_{2c}$ é a taxa de chegada de usuários do átomo 2;
- $\lambda_3 = \lambda_{3a} + \lambda_{3b} + \lambda_{3c}$ é a taxa de chegada de usuários do átomo 3;

Pode-se também escrever as seguintes relações entre as taxas de chegada dos tipos de subátomos (i.e., as classes de prioridades dos chamados):

- $\lambda_a = \lambda_{1a} + \lambda_{2a} + \lambda_{3a}$ é a taxa de chegada de usuários do sistema, com prioridade a ;
- $\lambda_b = \lambda_{1b} + \lambda_{2b} + \lambda_{3b}$ é a taxa de chegada de usuários do sistema, com prioridade b ;
- $\lambda_c = \lambda_{1c} + \lambda_{2c} + \lambda_{3c}$ é a taxa de chegada de usuários do sistema, com prioridade c ;

- $\lambda = \sum_{j=1}^{N_A} \lambda_j = \sum_{j=1}^{N_A} \sum_{k \in D} \lambda_{jk}$; a taxa total de chegada no sistema é a soma dos

N_A átomos geográficos, ou a soma de todos os $N_A \times |C|$ subátomos do sistema.

Pode-se notar que a subdivisão dos átomos em subátomos para incorporar

prioridades no modelo não invalida a hipótese 1 do modelo hipercubo, na qual a área deve ser dividida em átomos (agora em subátomos) geográficos. Essa abordagem já foi utilizada em Takeda *et. al.* (2007) para modelar as classes de usuários (avançadas e básicas) do SAMU-Campinas no modelo hipercubo (descrito no Capítulo 2), sem prejuízo da análise.

v. Admite-se que o tempo de serviço segue uma distribuição exponencial para cada servidor i com taxa de serviço μ_i , $i = 1, 2, 3$, de forma que a taxa de serviço

total do sistema seja $\mu = \sum_{i=1}^m \mu_i$.

vi. n é o número de usuários no sistema.

vii. A lista de preferências de despacho encontra-se na Tabela 4.1. Os servidores podem viajar a qualquer subátomo, sendo que para cada chamada, somente um servidor é enviado. Cada subátomo (1a, 2a, ..., 3c) possui um servidor primário e dois servidores *backup*. O servidor 1 (VSA) é enviado para atender a chamados de prioridade *b* e *c* somente se os servidores 2 e 3 (VSB's) estiverem ocupados. Assim, na lista de preferências da Tabela 4.1, o servidor 1 é o último a ser escolhido para ser enviado no atendimento a subátomos do tipo *b* e *c*, enquanto que nos subátomos do tipo *a* ele é o primeiro a ser escolhido. A Tabela 4.1 foi construída considerando o menor tempo médio de viagem de um servidor para cada subátomo, todos os tempos médios de viagem do servidor subátomo estão na Tabela 4.2.

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

Tabela 4.1 – Lista de preferências de despacho

	1a	1b	1c	2a	2b	2c	3a	3b	3c
1	10	10	10	12	12	12	15	15	15
2	13	13	13	10	10	10	14	14	14
3	10	10	10	12	12	12	15	15	15

Tabela 4.2 – Matriz do tempo de viagem do servidor i ao subátomo j , t_{ij} .

viii. O modelo utiliza disciplina de prioridade e admite fila com capacidade finita de, no máximo, três usuários em espera, como descrito na Seção 3.1.3. Caso ocorra um chamado quando a fila já tem três usuários esperando, este chamado será considerado uma perda para o sistema. Assim, a probabilidade de perda é uma medida de desempenho importante a ser considerada.

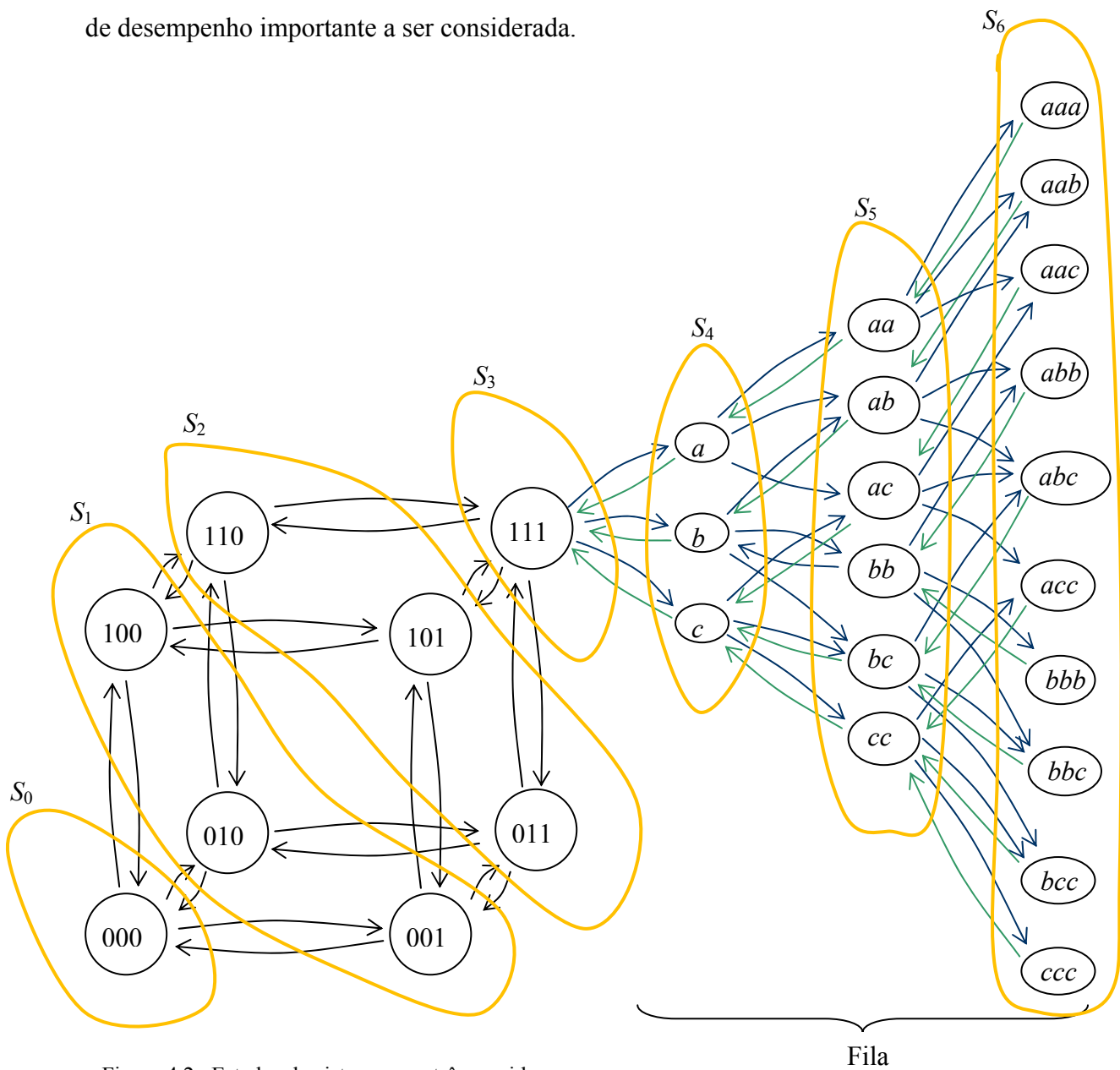


Figura 4.2– Estados do sistema com três servidores.

A Figura 4.2 representa todos os possíveis estados do exemplo ilustrativo. Em um sistema com três servidores, tem-se 2^3 estados possíveis, formando o hipercubo. Além disso, a cauda formada a partir do estado $B = \{111\}$, representa os estados do sistema utilizando disciplina de prioridade. Se o estado de um dado servidor i é representado por $b_i = \{0,1\}$, então um determinado estado do sistema B pode ser representado por $B = \{b_1, b_2, \dots, b_m\}$. Os estados do hipercubo correspondem aos vértices da Figura 4.2: $(\{000\}, \{001\}, \{010\}, \{100\}, \{011\}, \{101\}, \{110\}, \{111\})$. Considerando $r = 3$, o estado de algum usuário na fila pode ser representado por $s_j = \{a, b, c\}$. Assim, um estado da fila pode ser representado por $S_{C-n+1} = \{s_1 s_2 s_3\}$, sendo C a capacidade do sistema e n o número de usuários no sistema.

4.1.1 Transição de estados do modelo ilustrativo:

Na Figura 4.2, podemos observar que para passar de um estado a outro do sistema, corresponde a transitar de um vértice a outro adjacente. Isto ocorre com o término de um serviço (um servidor i passa de ocupado $b_i = 1$ para livre $b_i = 0$), ou com a chegada de uma chamada (um servidor passa de livre $b_i = 0$ para ocupado $b_i = 1$, lembrando que um único servidor é enviado para atender um chamado).

A seguir é discutido como determinar as equações de equilíbrio de alguns estados do sistema, todas as outras equações estão no Anexo G. Primeiramente, com até três usuários no sistema e, depois, com mais que três (fila com $n - 3$ usuários, a partir do estado $B = \{111\}$). A equação de equilíbrio do estado B do sistema pode ser obtida considerando as transições deste estado para os estados (vértices) adjacentes (fluxo para fora do estado) e as transições daqueles estados para o estado B (fluxo para dentro do estado), como discutido na Seção 2.4.

I. O estado $\{101\}$, com seus vértices adjacentes $\{111\}, \{100\}, \{001\}$, é mostrado na Figura 4.8. A equação de equilíbrio do estado $\{101\}$ é dada por:

$$(\lambda + \mu_3 + \mu_1)P_{\{101\}} = \mu_2 P_{\{111\}} + \lambda_1 P_{\{100\}} + \lambda_a P_{\{001\}} \quad (4.1)$$

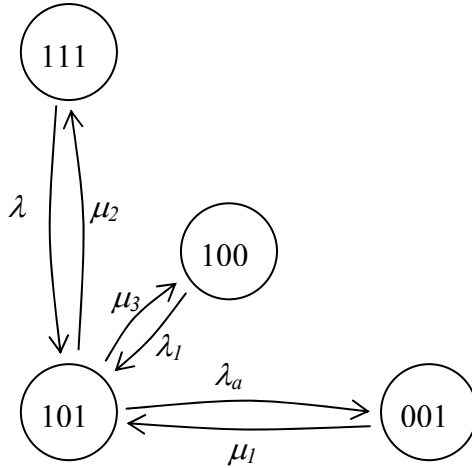


Figura 4.3 – Vértice $\{101\}$ e seus adjacentes.

O lado esquerdo da Equação (4.6) corresponde ao fluxo para fora do estado $\{101\}$. Na Tabela 4.1 e na Figura 4.8, as transições de estado para os vértices adjacentes a $\{101\}$ são:

- $\{101\} \rightarrow \{111\}$, quando ocorre uma chamada em qualquer subátomos (preferencial ou *backup*) de qualquer prioridade, com taxa λ ;
- $\{101\} \rightarrow \{100\}$, quando ocorre um término de serviço do servidor 1, com taxa μ_1 ;
- $\{101\} \rightarrow \{001\}$, quando ocorre um término de serviço do servidor 3, com taxa μ_3 .

O lado direito da Equação (4.6) corresponde ao fluxo para dentro do estado $\{101\}$. Na Tabela 4.1 e na Figura 4.8, as transições de estado para os vértices adjacentes a $\{101\}$ são:

- $\{111\} \rightarrow \{101\}$, quando ocorre um término de serviço do servidor 2, com taxa μ_2 ;
- $\{100\} \rightarrow \{101\}$, quando ocorre uma chamada nos subátomos $1a$, $1b$ ou $1c$, com taxa λ_1 ;

➤ $\{001\} \rightarrow \{101\}$, quando ocorre uma chamada nos subátomos $1a$, $2a$ ou $3a$, com taxa λ_a .

II. O estado $\{011\}$, com seus vértices adjacentes $\{111\}, \{001\}, \{010\}$, é mostrado na Figura 4.9. A equação de equilíbrio do estado $\{011\}$ é dada por:

$$(\lambda + \mu_2 + \mu_3)P_{\{011\}} = (\lambda_b + \lambda_c)P_{\{010\}} + (\lambda_b + \lambda_c)P_{\{001\}} + \mu_1 P_{\{111\}} \quad (4.2)$$

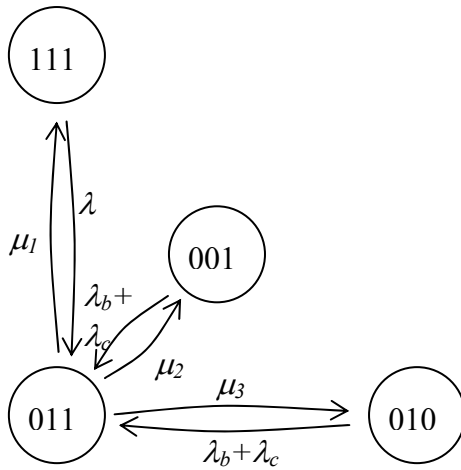


Figura 4.4 – Vértice $\{011\}$ e seus adjacentes.

O lado esquerdo da Equação (4.7) corresponde ao fluxo para fora do estado $B = \{011\}$. Na Tabela 4.1 e na Figura 4.9, as transições de estado para os vértices adjacentes a $\{011\}$ são:

- $\{011\} \rightarrow \{111\}$, quando ocorre uma chamada em qualquer, com taxa λ ;
- $\{011\} \rightarrow \{001\}$, quando ocorre um término de serviço do servidor 2, com taxa μ_2 ;
- $\{011\} \rightarrow \{010\}$, quando ocorre um término de serviço do servidor 3, com taxa μ_3 .

O lado direito da Equação (4.7) corresponde ao fluxo para dentro do estado $B = \{011\}$. Na Tabela 4.1 e na Figura 4.9, as transições de estado para os vértices adjacentes a $\{011\}$ são:

- $\{111\} \rightarrow \{011\}$, quando ocorre um término de serviço do servidor 1, com taxa μ_1 ;
- $\{001\} \rightarrow \{011\}$, quando ocorre uma chamada nos subátomos $1b$, $1c$, $2b$, $2c$, $3b$ ou $3c$, com taxa $(\lambda_b + \lambda_c)$;
- $\{010\} \rightarrow \{011\}$, quando ocorre uma chamada nos subátomos $1b$, $1c$, $2b$, $2c$, $3b$ ou $3c$, com taxa $(\lambda_b + \lambda_c)$.

O sistema de equações de equilíbrio com todos os oito estados $\{000\}, \{001\}, \{010\}, \{100\}, \{011\}, \{101\}, \{110\}, \{111\}$ (lembrando que $S_4 = \{a, b, c\}$), é:

$$\begin{aligned}
 \{000\} &\rightarrow (\lambda) P_{\{000\}} = \mu_3 P_{\{001\}} + \mu_2 P_{\{010\}} + \mu_1 P_{\{100\}} \\
 \{100\} &\rightarrow (\lambda + \mu_1) P_{\{100\}} = \mu_2 P_{\{110\}} + \mu_3 P_{\{101\}} + \lambda_a P_{\{000\}} \\
 \{010\} &\rightarrow (\lambda + \mu_2) P_{\{010\}} = \mu_1 P_{\{110\}} + \mu_3 P_{\{011\}} + (\lambda_{2b} + \lambda_{2c} + \lambda_{3b} + \lambda_{3c}) P_{\{000\}} \\
 \{001\} &\rightarrow (\lambda + \mu_3) P_{\{001\}} = (\lambda_{1b} + \lambda_{1c}) P_{\{000\}} + \mu_1 P_{\{101\}} + \mu_2 P_{\{011\}} \\
 \{110\} &\rightarrow (\lambda + \mu_1 + \mu_2) P_{\{110\}} = \mu_3 P_{\{111\}} + \lambda_a P_{\{010\}} + (\lambda_2 + \lambda_3) P_{\{100\}} \\
 \{101\} &\rightarrow (\lambda + \mu_3 + \mu_1) P_{\{101\}} = \mu_2 P_{\{111\}} + \lambda_1 P_{\{100\}} + \lambda_a P_{\{001\}} \\
 \{011\} &\rightarrow (\lambda + \mu_2 + \mu_3) P_{\{011\}} = (\lambda_b + \lambda_c) P_{\{010\}} + (\lambda_b + \lambda_c) P_{\{001\}} + \mu_1 P_{\{111\}} \\
 \{111\} &\rightarrow (\lambda + \mu) P_{\{111\}} = \lambda P_{\{011\}} + \lambda P_{\{101\}} + \lambda P_{\{110\}} + \mu (P_{\{a\}} + P_{\{b\}} + P_{\{c\}})
 \end{aligned} \tag{4.3}$$

Apenas a última equação é diferente do modelo hipercubo original descrito na Seção 3.2, em todas as outras o sistema sai de seu estado corrente para outro com a chegada de uma chamada de qualquer subátomo, com taxa total λ . Além disso, uma chamada pode entrar em fila, caso todos os servidores estejam ocupados (estado $\{111\}$).

Conforme descrito na Seção 3.5, Chiyoshi *et al* (2000) mencionam que o sistema acima escrito na forma matricial $Ax = b$ é um sistema linear homogêneo indeterminado. Isto ocorre porque as equações apenas impõem condições de equilíbrio para cada estado do sistema, mas nada especifica sobre como a massa total de probabilidade se distribui entre estes estados e os estados da cauda. Uma maneira de tornar o sistema possível e determinado é a substituição de uma das equações do sistema

por uma equação de normalização, considerando que $\sum_{n=0}^N P_B = 1$ (a soma de todos os estados do sistema deve ser igual a 1), sendo que N é o número de estados possíveis para o sistema com até três usuários em fila. A equação de normalização é dada por:

$$P_{\{000\}} + P_{\{001\}} + P_{\{010\}} + P_{\{100\}} + \dots + P_{\{111\}} + P_{\{a\}} + \dots + P_{\{ccc\}} = 1 \quad (4.4)$$

A condição de equilíbrio do sistema requer que $(\lambda)P_{\{111\}} = \mu P_{\{S_4\}}$, como $P_{\{S_4\}} = P_{\{a\}} + P_{\{b\}} + P_{\{c\}}$, temos que: $(\lambda)P_{\{111\}} = \mu(P_{\{a\}} + P_{\{b\}} + P_{\{c\}})$ e $\frac{\lambda}{\mu} = \rho$. Em que $\rho < 1$, temos a equação simplificada para:

$$\triangleright P_{\{a\}} + P_{\{b\}} + P_{\{c\}} = \frac{\lambda}{\mu} P_{\{111\}}.$$

De forma similar, para a transição dos estados correspondentes a dois e três usuários na fila $S_5 = \{aa, ab, ac, bb, bc, cc\}$ e $S_6 = \{aaa, aab, aac, \dots, ccc\}$. Tem-se que

$P_{\{S_5\}} = P_{\{aa\}} + P_{\{ab\}} + P_{\{ac\}} + P_{\{bb\}} + P_{\{bc\}} + P_{\{cc\}}$ e $P_{\{S_6\}} = P_{\{aaa\}} + P_{\{aab\}} + \dots + P_{\{ccc\}}$. Dessa forma:

$$\triangleright P_{\{aa\}} + P_{\{ab\}} + P_{\{ac\}} + P_{\{bb\}} + P_{\{bc\}} + P_{\{cc\}} = \left(\frac{\lambda}{\mu}\right)^2 P_{\{111\}}; \text{ e}$$

$$\triangleright P_{\{aaa\}} + P_{\{aab\}} + P_{\{aac\}} + P_{\{abb\}} + P_{\{abc\}} + P_{\{acc\}} + P_{\{bbb\}} + P_{\{bbc\}} + P_{\{bcc\}} + P_{\{ccc\}} = \left(\frac{\lambda}{\mu}\right)^3 P_{\{111\}}$$

Assim, somando as probabilidades dos estados nos quais todos os servidores estão ocupados, tem-se:

$$P_{\{111\}} + P_{\{a\}} + \dots + P_{\{ccc\}} = P_{\{111\}} + \rho P_{\{111\}} + \rho^2 P_{\{111\}} + \rho^3 P_{\{111\}} = P_{\{111\}} \sum_{j=0}^{n-N} \rho^j. \quad (4.5)$$

Pode-se notar que, como $\rho < 1$, $\sum_{j=0}^{n-N} \rho^j = \frac{(1 - \rho^{n-N+1})}{(1 - \rho)}$ (uma série geométrica

finita de razão ρ). Portanto,

$$P_{\{111\}} + P_{\{a\}} + \dots + P_{\{ccc\}} = \frac{(1 - \rho^{n-N+1}) P_{\{111\}}}{(1 - \rho)}. \quad (4.6)$$

Substituindo na equação de normalização, tem-se:

$$P_{\{000\}} + P_{\{100\}} + P_{\{010\}} + P_{\{001\}} + \dots + P_{\{011\}} + \frac{(1 - \rho^{n-N+1}) P_{\{111\}}}{(1 - \rho)} = 1$$

Na Figura 4.11 a seguir, pode-se observar o espaço de estados da fila S .

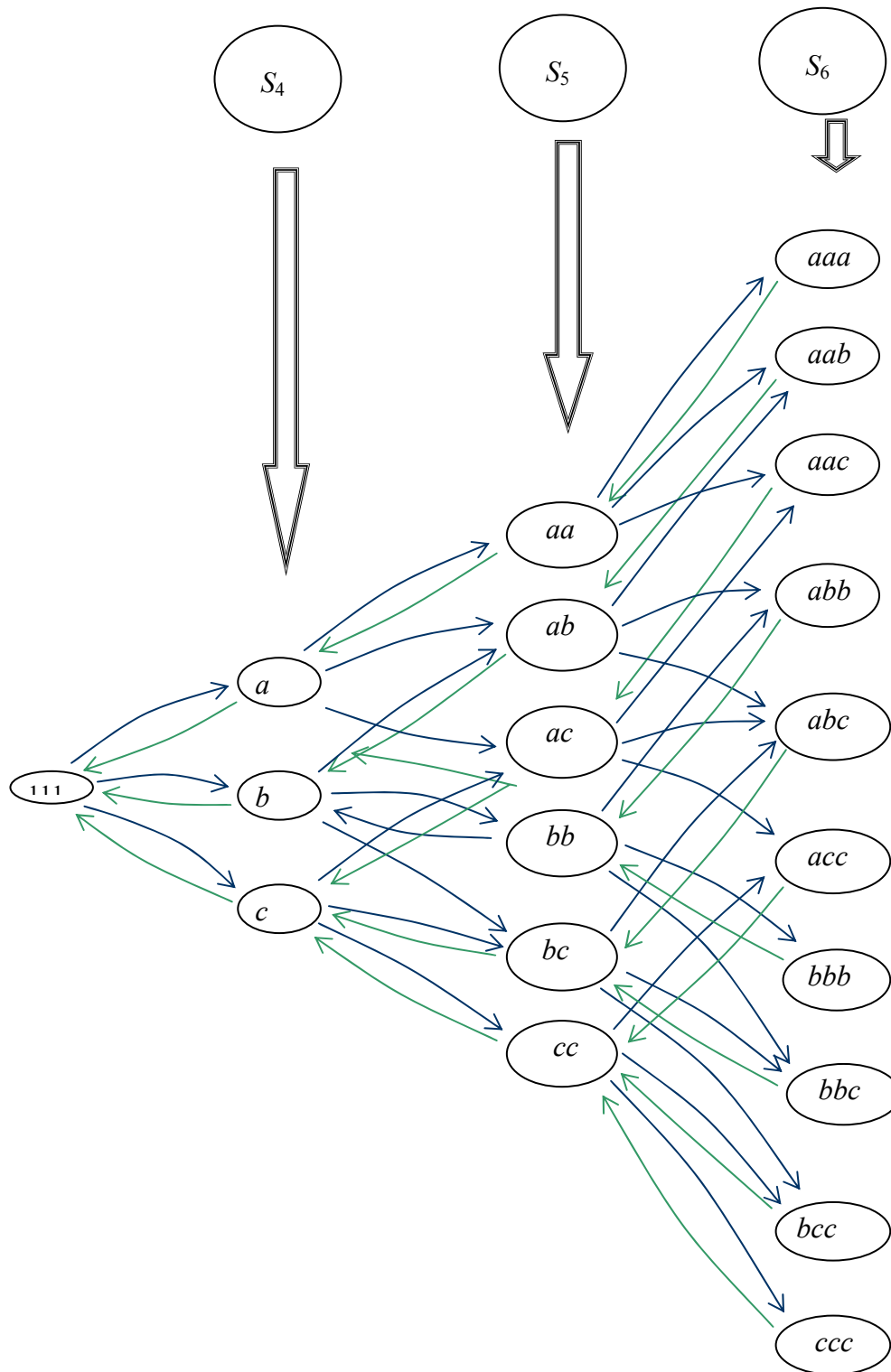


Figura 4.5 – Espaço de estados da fila 3 classes de usuários e com prioridades.

Como pode ser observado na Figura 4.11, uma fila se forma a partir do estado $\{111\}$, na qual temos até três usuários em fila. Esta fila é formada a partir dos chamados com qualquer prioridade, vindos de qualquer subátomo (fila S). É importante notar que o número de estados da fila não depende de m (número de servidores), de tal forma que tanto para um sistema com 3 ou 10 servidores o número de estados da fila é o mesmo. O número de estados da fila depende apenas da capacidade do sistema.

De forma geral, pode-se encontrar o número de estados para r classes de prioridade e n usuários na fila (combinação de r objetos n a n , com repetição):

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

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

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

Para o caso particular aonde $n = 3$ e $r = 3$, tem-se que:

$$\begin{aligned} f(3, 1) &= \binom{3+1-1}{1} = 3; \\ f(3, 2) &= \binom{3+2-1}{2} = 6; \\ f(3, 3) &= \binom{3+3-1}{3} = 10. \end{aligned}$$

Assim, o número de estados para um sistema com capacidade para 3 usuários em fila, trabalhando com 3 classes de usuários é dado por:

$$N = \sum_{k=1}^3 f(r, n) = 3 + 6 + 10 = 19 \text{ estados. O número total de estados no sistema é}$$

$$2^m + \binom{r+L}{L} - 1.$$

4.1.2 Transição de estados da fila S :

A Figura 4.11 mostra a transição de estados da fila S utilizando uma disciplina de prioridades, de forma que o chamado com maior prioridade é atendido primeiro. A seguir é feita a análise e construção das equações de equilíbrio da fila do sistema a partir do estado $\{111\}$. Tem-se dezenove possíveis estados para o sistema $\{a\}, \{b\}, \{c\}, \{aa\}, \{ab\}, \{ac\}, \{bb\}, \{bc\}, \{cc\}, \{aaa\}, \{aab\}, \{aac\}, \{abb\}, \{abc\}, \{acc\}, \{bbb\}, \{bbc\}, \{bcc\}, \{ccc\}$. Dois estados são analisados a seguir, a análise dos outros estados está no Anexo H.

I. O estado $\{ab\}$, com seus vértices adjacentes $\{a\}, \{b\}, \{abb\}, \{aab\}, \{abc\}$, é mostrado na Figura 4.16. A equação de equilíbrio do estado $\{ab\}$ é dada por:

$$(\lambda + \mu)P_{\{ab\}} = \lambda_a P_{\{b\}} + \lambda_b P_{\{a\}} + \mu P_{\{aab\}} \quad (4.7)$$

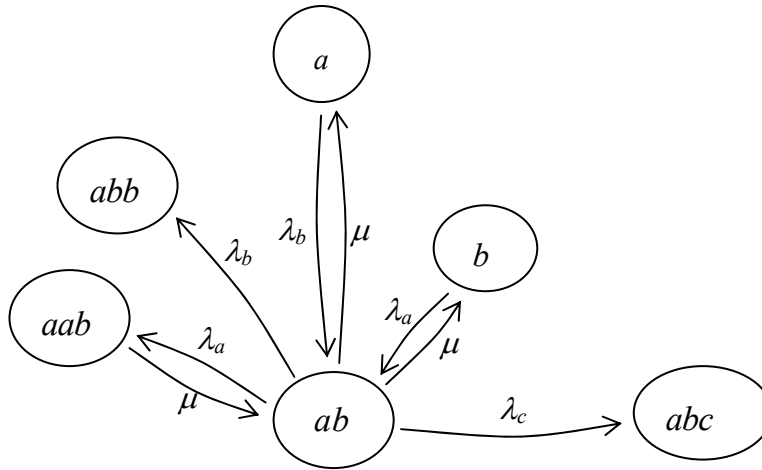


Figura 4.6— Vértice $\{ab\}$ e seus adjacentes.

O lado esquerdo da Equação (4.17) corresponde ao fluxo para fora do estado $B = \{ab\}$. Na Equação (4.17) e na Figura 4.16, as transições de estado para os vértices adjacentes a $\{ab\}$ são:

- $\{ab\} \rightarrow \{b\}$, quando ocorre um término de serviço, com taxa μ ;
- $\{ab\} \rightarrow \{aab\}$, quando ocorre uma chamada nos subátomos $1a$, $2a$ ou $3a$,

com taxa λ_a ;

➤ $\{ab\} \rightarrow \{abb\}$, quando ocorre uma chamada nos subátomos $1b, 2b$ ou $3b$, com taxa λ_b ;

➤ $\{ab\} \rightarrow \{abc\}$, quando ocorre uma chamada nos subátomos $1c, 2c$, ou $3c$, com taxa λ_c .

O lado direito da Equação (4.17) corresponde ao fluxo para dentro do estado $B = \{ab\}$. Na Equação (4.17) e na Figura 4.16, as transições de estado para os vértices adjacentes a $\{ab\}$ são:

➤ $\{a\} \rightarrow \{ab\}$, quando ocorre uma chamada nos subátomos $1b, 2b$ ou $3b$, com taxa λ_b ;

➤ $\{b\} \rightarrow \{ab\}$, quando ocorre uma chamada nos subátomos $1a, 2a$ ou $3a$, com taxa λ_a ;

➤ $\{aab\} \rightarrow \{ab\}$, quando ocorre um término de serviço, com taxa μ .

II. O estado $\{ac\}$, com seus vértices adjacentes $\{a\}, \{c\}, \{acc\}, \{aac\}, \{abc\}$, é mostrado na Figura 4.17. A equação de equilíbrio do estado $\{ac\}$ é dada por:

$$(\lambda + \mu)P_{\{ac\}} = \lambda_c P_{\{a\}} + \lambda_a P_{\{c\}} + \mu P_{\{aac\}} \quad (4.8)$$

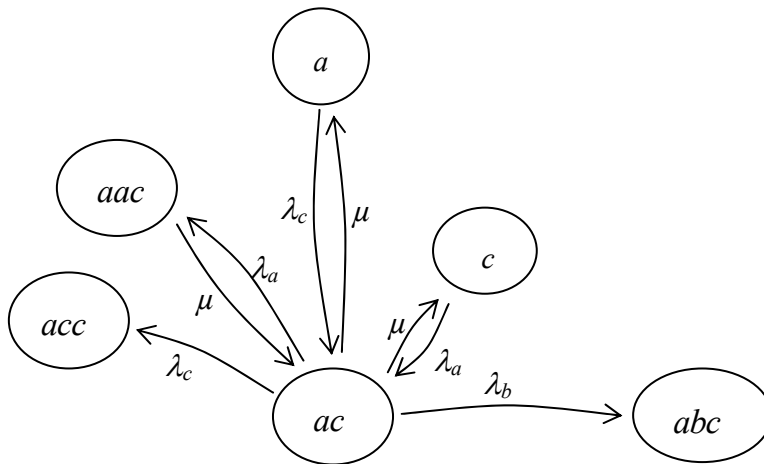


Figura 4.7 – Vértice $\{ac\}$ e seus adjacentes.

O lado esquerdo da Equação (4.18) corresponde ao fluxo para fora do estado $B = \{ac\}$. Na Equação (4.18) e na Figura 4.17, as transições de estado para os vértices adjacentes a $\{ac\}$ são:

- $\{ac\} \rightarrow \{c\}$, quando ocorre um término de serviço, com taxa μ ;
- $\{ac\} \rightarrow \{aac\}$, quando ocorre uma chamada nos átomos $1a$, $2a$ ou $3a$, com taxa λ_a ;
- $\{ac\} \rightarrow \{acc\}$, quando ocorre uma chamada nos átomos $1c$, $2c$, ou $3c$, com taxa λ_c ;
- $\{ac\} \rightarrow \{abc\}$, quando ocorre uma chamada nos átomos $1b$, $2b$ ou $3b$, com taxa λ_b .

O lado direito da Equação (4.18) corresponde ao fluxo para dentro do estado $B = \{ac\}$. Na Equação (4.18) e na Figura 4.17, as transições de estado para os vértices adjacentes a $\{ac\}$ são:

- $\{a\} \rightarrow \{ac\}$, quando ocorre uma chamada nos átomos $1c$, $2c$, ou $3c$, com taxa λ_c ;
- $\{c\} \rightarrow \{ac\}$, quando ocorre uma chamada nos átomos $1a$, $2a$ ou $3a$, com taxa λ_a ;
- $\{aac\} \rightarrow \{ac\}$, quando ocorre um término de serviço, com taxa μ .

Assim, o sistema de equações que deriva as probabilidades de estado da fila S é:

$$\{a\} \rightarrow (\lambda + \mu)P_{\{a\}} = \mu P_{\{aa\}} + \lambda_a P_{\{111\}} \quad (4.9)$$

$$\{b\} \rightarrow (\lambda + \mu)P_{\{b\}} = \lambda_b P_{\{111\}} + \mu P_{\{ab\}} + \mu P_{\{bb\}}$$

$$\{c\} \rightarrow (\lambda + \mu)P_{\{c\}} = \lambda_c P_{\{111\}} + \mu P_{\{cc\}} + \mu P_{\{ac\}} + \mu P_{\{bc\}}$$

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

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

$$\{ac\} \rightarrow (\lambda + \mu)P_{\{ac\}} = \lambda_c P_{\{a\}} + \lambda_a P_{\{c\}} + \mu P_{\{aac\}}$$

$$\begin{aligned}
\{bc\} &\rightarrow (\lambda + \mu)P_{\{bc\}} = \lambda_c P_{\{b\}} + \lambda_b P_{\{c\}} + \mu P_{\{abc\}} + \mu P_{\{bbc\}} \\
\{bb\} &\rightarrow (\lambda + \mu)P_{\{bb\}} = \lambda_b P_{\{b\}} + \mu P_{\{abb\}} + \mu P_{\{bbb\}} \\
\{cc\} &\rightarrow (\lambda + \mu)P_{\{cc\}} = \lambda_c P_{\{c\}} + \mu P_{\{bcc\}} + \mu P_{\{acc\}} + \mu P_{\{ccc\}} \\
\{aaa\} &\rightarrow \mu P_{\{aaa\}} = \lambda_a P_{\{aa\}} \\
\{aab\} &\rightarrow \mu P_{\{aab\}} = \lambda_a P_{\{ab\}} + \lambda_b P_{\{aa\}} \\
\{aac\} &\rightarrow \mu P_{\{aac\}} = \lambda_a P_{\{ac\}} + \lambda_c P_{\{aa\}} \\
\{abb\} &\rightarrow \mu P_{\{abb\}} = \lambda_b P_{\{ab\}} + \lambda_a P_{\{bb\}} \\
\{abc\} &\rightarrow \mu P_{\{abc\}} = \lambda_a P_{\{bc\}} + \lambda_b P_{\{ac\}} + \lambda_c P_{\{ab\}} \\
\{acc\} &\rightarrow \mu P_{\{acc\}} = \lambda_a P_{\{cc\}} + \lambda_c P_{\{ac\}} \\
\{bbc\} &\rightarrow \mu P_{\{bbc\}} = \lambda_b P_{\{bc\}} + \lambda_c P_{\{bb\}} \\
\{bbb\} &\rightarrow \mu P_{\{bbb\}} = \lambda_b P_{\{bb\}} \\
\{bcc\} &\rightarrow \mu P_{\{bcc\}} = \lambda_c P_{\{bc\}} + \lambda_b P_{\{cc\}} \\
\{ccc\} &\rightarrow \mu P_{\{ccc\}} = \lambda_c P_{\{cc\}}
\end{aligned}$$

Pode-se notar que o sistema de equações (4.32) tem como parâmetro $P_{\{111\}}$, uma vez que o modelo hipercubo pode ser resolvido separadamente. Assim, este sistema pode ser resolvido com a informação da probabilidade de todos os servidores estarem ocupados, obtido pela resolução do sistema (4.9). No final deste capítulo (Seção 4.4) apresenta-se um procedimento para generalizar a geração deste sistema de equações, para qualquer número r de classes de prioridades (ao invés de apenas $r = 3$) e para qualquer tamanho máximo de fila $C-m$ (ao invés de apenas $C-m = 3$, i.e., S_1 , S_2 e S_3).

4.2 Medidas de Desempenho do Exemplo Ilustrativo

A Seção 3.5 mostra as medidas de desempenho usuais utilizadas no modelo hipercubo clássico. Nesta seção, são apresentadas todas as medidas de desempenho utilizadas no modelo hipercubo com prioridade na fila, com as devidas modificações realizadas sobre as expressões do modelo hipercubo clássico.

Workload

O *workload* é a fração de tempo em que o servidor permanece ocupado, o qual é calculado pela soma das probabilidades de o mesmo estar ocupado. Pode ser

obtida pela expressão:

$$\rho_i = \sum_{\{B:b_i=1\}} P_B + P_Q, \text{ em que:}$$

- ρ_i é a carga de trabalho (*workload*) do servidor i ($i = 1, 2, \dots, m$);
- $\sum_{\{B:b_i=1\}} P_B$ é a soma das probabilidades dos estados (de $\{000\}$ a $\{111\}$), em

que o servidor i está ocupado ($b_i = 1$);

- P_Q é a probabilidade de fila do sistema.

Para o exemplo ilustrativo acima, tem-se:

$$\begin{aligned} \rho_1 &= P_{\{100\}} + P_{\{110\}} + P_{\{101\}} + P_{\{111\}} + P_Q \\ \rho_2 &= P_{\{010\}} + P_{\{110\}} + P_{\{011\}} + P_{\{111\}} + P_Q \\ \rho_3 &= P_{\{001\}} + P_{\{011\}} + P_{\{101\}} + P_{\{111\}} + P_Q, \text{ em que:} \end{aligned} \quad (4.10)$$

$$P_Q = P_{\{a\}} + \dots + P_{\{ccc\}} = 1 - (P_{\{000\}} + P_{\{001\}} + \dots + P_{\{111\}}).$$

Frequências de despacho

A frequência de despacho é a fração de todos os despachos do servidor i ao átomo (ou subátomo) j (f_{ij}) e pode ser decomposta em duas partes: $f_{ij}^{[nq]}$, que corresponde à fração de despachos de um servidor i para um átomo j que não implica tempo de espera em fila para o usuário; $f_{ij}^{[q]}$, que corresponde à fração de despachos de um servidor i para um átomo j sujeito a espera em fila.

$$f_{ij} = f_{ij}^{[nq]} + f_{ij}^{[q]} = \underbrace{\frac{\lambda_j}{\lambda} \sum_{B \in E_{ij}} P_B}_{f_{ij}^{[nq]}} + \underbrace{\frac{\lambda_j}{\lambda} P_Q \frac{\mu_i}{\mu}}_{f_{ij}^{[q]}}. \text{ Em que,}$$

E_{ij} é o conjunto dos estados nos quais o servidor i é o primeiro servidor disponível na lista de despacho do átomo j .

$$P_{Q'} \text{ é a probabilidade de saturação do sistema } (P_{Q'} = P_Q + P_{\{111\}}).$$

P_Q é a probabilidade de saturação do sistema menos a probabilidade de todos os servidores estarem ocupados, isto é, a probabilidade de fila do sistema.

No exemplo ilustrativo, considera-se três servidores e três átomos (com três subátomos cada um). Além de fila com três tipos de prioridades, as quais podem ser atendidas por qualquer servidor. Dessa forma, tem-se que:

$$f_{ij} = \frac{\lambda_j}{\lambda} \sum_{B \in E_{ij}} P_B + \frac{\lambda_j}{\lambda} P_{\{111\}} \frac{\mu_i}{\mu} + \frac{\lambda_j}{\lambda} P_Q \frac{\mu_i}{\mu},$$

$$f_{ij} = \frac{\lambda_{ja} + \lambda_{jb} + \lambda_{jc}}{\lambda} \sum_{B \in E_{ij}} P_B + \frac{\lambda_{ja} + \lambda_{jb} + \lambda_{jc}}{\lambda} P_{\{111\}} \frac{\mu_i}{\mu} + \frac{\lambda_j}{\lambda} (P_{Qa} + P_{Qb} + P_{Qc}) \frac{\mu_i}{\mu}.$$

Em que:

$$\lambda_j = \lambda_{ja} + \lambda_{jb} + \lambda_{jc}$$

$$P_Q = P_{Qa} + P_{Qb} + P_{Qc}.$$

Considerando que $f_{ij} = f_{ija} + f_{ijb} + f_{ijc}$, tem-se:

$$f_{ija} = \frac{\lambda_{ja}}{\lambda} \sum_{B \in E_{ij}} P_B + \frac{\lambda_{ja}}{\lambda} P_{\{111\}} \frac{\mu_i}{\mu} + \frac{\lambda_j}{\lambda} P_{Qa} \frac{\mu_i}{\mu}, \quad (4.11)$$

$$f_{ijb} = \frac{\lambda_{jb}}{\lambda} \sum_{B \in E_{ij}} P_B + \frac{\lambda_{jb}}{\lambda} P_{\{111\}} \frac{\mu_i}{\mu} + \frac{\lambda_j}{\lambda} P_{Qb} \frac{\mu_i}{\mu},$$

$$f_{ijc} = \frac{\lambda_{jc}}{\lambda} \sum_{B \in E_{ij}} P_B + \frac{\lambda_{jc}}{\lambda} P_{\{111\}} \frac{\mu_i}{\mu} + \frac{\lambda_j}{\lambda} P_{Qc} \frac{\mu_i}{\mu} \text{ em que,}$$

E_{ija} é o conjunto de estados em que o servidor i é o primeiro servidor disponível na lista de despacho do subátomo ja .

E_{ijb} é o conjunto de estados em que o servidor i é o primeiro servidor disponível na lista de despacho do subátomo jb .

E_{ijc} são o conjunto de estados em que o servidor i é o primeiro servidor disponível na lista de despacho do subátomo jc .

P_{Qa} é a soma dos estados da fila onde há pelo menos um chamado da classe a .

P_{Qb} é a soma dos estados da fila onde há pelo menos um chamado da classe b , mas não tem chamados do tipo a .

P_{Qc} é a soma dos estados da fila onde há pelo menos um chamado da classe c , mas não tem chamados do tipo a e b .

$$\mu = \sum_{i=1}^m \mu_i ,$$

$P_{\{111\}}$ é a probabilidade de todos os três servidores estarem ocupados,

$\frac{\mu_i}{\mu}$ indica a probabilidade de o servidor i ser o primeiro a terminar o serviço, considerando variáveis aleatórias independentes e exponencialmente distribuídas.

As frequências de despacho são calculadas a partir da distribuição de equilíbrio obtidas a partir do modelo hipercubo com prioridade. Assim, as medidas de desempenho a seguir são reescritas de forma a considerar sempre as frequências de despacho.

Tempos médios de viagem

Os tempos médios de viagem são obtidos a partir da matriz origem-destino dos tempos de viagem entre todos os pares de subátomos, $\tau_{pl,jk}$ em que $l \in D$, $k \in D$, $p = 1, 2, \dots, N_A$ e $j = 1, 2, \dots, N_A$. Eles geralmente refletem a influência de fatores como tráfego, horário, presença de barreiras. Assim, o tempo gasto para um servidor ir do subátomo pl ao subátomo jk não necessariamente coincide com o tempo em que o servidor gasta para ir do subátomo jk ao subátomo pl , ou seja, nem sempre $\tau_{pl,jk} = \tau_{jk,pl}$ (LARSON e ODoni, 2007).

Tempo médio de viagem no sistema

Para calcular o tempo médio de viagem para o sistema é necessário conhecer a localização dos servidores, o tempo médio necessário para um servidor m , quando disponível, viajar até o subátomo jk , e o tempo médio de espera de um chamado que está em fila.

A representação da localização dos servidores é feita a partir de uma matriz $L = [l_{i,jk}]$, em que os elementos representam a probabilidade de um servidor i estar localizado em um subátomo jk , quando disponível. Sendo L uma matriz estocástica, ou seja $\sum_{j=1}^{N_A} \sum_{k \in D} l_{i,jk} = 1$, se o servidor i está localizado no subátomo jk , então $l_{i,jk} = 1$, e $l_{i,jk} = 0$ se o servidor i não está localizado no átomo jk . Assim, o tempo médio de viagem para um servidor se deslocar até um determinado subátomo é dado por:

$$t_{i,pl} = \sum_{j=1}^{N_A} \sum_{k \in D} l_{i,jk} \cdot \tau_{jk,pl}.$$

O tempo médio de viagem para chamados em fila, com a disciplina FCFS em um sistema em que não há prioridade, é dado por:

$$\bar{T}_Q = \sum_{p=1}^{N_A} \sum_{l \in D} \sum_{j=1}^{N_A} \sum_{k \in D} \frac{\lambda_{pl} \lambda_{jk}}{\lambda^2} \tau_{pl,jk}.$$

em que as razões $\frac{\lambda_{pl}}{\lambda}$ e $\frac{\lambda_{jk}}{\lambda}$ correspondem à probabilidade de um chamado

que está em fila ter sido gerado no subátomo pl , e à probabilidade deste chamado ser atendido por um servidor localizado no subátomo jk , respectivamente. Nota-se que a expressão acima considera que o servidor irá do subátomo pl diretamente para o subátomo jk (consumindo tempo $\tau_{pl,jk}$), sem ter que voltar para sua base antes de atender no subátomo jk .

Assim, o tempo médio de viagem no sistema para fila FCFS, considerando subátomos, é dado por:

$$\bar{T} = \sum_{i=1}^m \sum_{j=1}^{N_A} \sum_{k \in D} f_{i,jk}^{[nq]} t_{i,jk}^{[nq]} + P_{Q'} \bar{T}_Q, \quad (4.12)$$

A Equação 4.35 (LARSON, 1974) é reescrita a seguir, para incorporar disciplina de prioridade na fila e subátomos geográficos, admitindo servidores homogêneos:

$$\begin{aligned} \bar{T} &= \sum_{i=1}^m \sum_{j=1}^{N_A} \sum_{k \in D} f_{i,jk}^{[nq]} t_{i,jk}^{[nq]} + \sum_{i=1}^m \sum_{j=1}^{N_A} \sum_{k \in D} f_{i,jk}^{[q]} t_{i,jk}^{[q]}, \text{ em que} \\ \sum_{i=1}^m \sum_{j=1}^{N_A} \sum_{k \in D} f_{i,jk}^{[q]} t_{i,jk}^{[q]} &= \underbrace{\sum_{i=1}^m \sum_{j=1}^{N_A} \sum_{k \in D} \frac{\lambda_{jk}}{\lambda} P_{Q'}}_{f_{i,jk}^{[q]}} \underbrace{\frac{1}{m} \sum_{p=1}^{N_A} \sum_{l \in D} \frac{\lambda_{pl}}{\lambda} \tau_{pl,jk}}_{t_{i,jk}^{[q]}} \\ &= P_{Q'} \frac{1}{m} \sum_{i=1}^m \sum_{j=1}^{N_A} \sum_{k \in D} \frac{\lambda_{jk}}{\lambda} \sum_{p=1}^{N_A} \sum_{l \in D} \frac{\lambda_{pl}}{\lambda} \tau_{pl,jk}, \\ &= P_{Q'} \sum_{j=1}^{N_A} \sum_{k \in D} \frac{\lambda_{jk}}{\lambda} \sum_{p=1}^{N_A} \sum_{l \in D} \frac{\lambda_{pl}}{\lambda} \tau_{pl,jk} \sum_{i=1}^m \frac{1}{m}, \text{ como } \sum_{i=1}^m \frac{1}{m} = 1, \text{ tem-se:} \\ &= P_{Q'} \sum_{j=1}^{N_A} \sum_{k \in D} \sum_{p=1}^{N_A} \sum_{l \in D} \frac{\lambda_{jk}}{\lambda} \frac{\lambda_{pl}}{\lambda} \tau_{pl,jk}. \end{aligned}$$

$f_{ij}^{[q]}$ corresponde à proporção de chamados que foram gerados no átomo j , sujeito a tempo de espera em fila.

No caso de servidores heterogêneos, como é o exemplo ilustrativo descrito na Seção 4.1, tem-se que:

$$\begin{aligned} \bar{T} &= \sum_{i=1}^m \sum_{j=1}^{N_A} \sum_{k \in D} f_{i,jk}^{[nq]} t_{i,jk}^{[nq]} + \sum_{i=1}^m \sum_{j=1}^{N_A} \sum_{k \in D} f_{i,jk}^{[q]} t_{i,jk}^{[q]}, \text{ em que} \quad (4.13) \\ \sum_{i=1}^m \sum_{j=1}^{N_A} \sum_{k \in D} f_{i,jk}^{[q]} t_{i,jk}^{[q]} &= \underbrace{\sum_{i=1}^m \sum_{j=1}^{N_A} \sum_{k \in D} \frac{\lambda_{jk}}{\lambda} P_{Q'}}_{f_{i,jk}^{[q]}} \underbrace{\frac{\mu_i}{\mu} \sum_{p=1}^{N_A} \sum_{l \in D} \frac{\lambda_{pl}}{\lambda} \tau_{pl,jk}}_{t_{i,jk}^{[q]}} \\ &= P_{Q'} \sum_{j=1}^{N_A} \sum_{k \in D} \frac{\lambda_{jk}}{\lambda} \sum_{p=1}^{N_A} \sum_{l \in D} \frac{\lambda_{pl}}{\lambda} \tau_{pl,jk} \sum_{i=1}^m \frac{\mu_i}{\mu}, \text{ como } \sum_{i=1}^m \frac{\mu_i}{\mu} = 1, \text{ tem-se:} \\ &= P_{Q'} \sum_{j=1}^{N_A} \sum_{k \in D} \sum_{p=1}^{N_A} \sum_{l \in D} \frac{\lambda_{jk}}{\lambda} \frac{\lambda_{pl}}{\lambda} \tau_{pl,jk}. \end{aligned}$$

Assim, tanto para o caso de servidores homogêneos ou heterogêneos, a

expressão do tempo médio de viagem no sistema é a mesma.

Tempo médio de viagem ao subátomo

O tempo médio de viagem para cada subátomo jk (\bar{T}_{jk}) é outra medida que reflete o nível de serviço oferecido pelo sistema. Como descrito na Seção 3.5, a equação do modelo hipercubo clássico, escrita a seguir, com as devidas adaptações de átomo para subátomo, dá o tempo médio de viagem ao subátomo jk utilizando a disciplina FCFS.

$$\bar{T}_{jk} = \frac{\sum_{i=1}^m f_{i,jk}^{[nq]} t_{i,jk}^{[nq]}}{\sum_{i=1}^m f_{i,jk}^{[nq]}} (1 - P_{Q'}) + \sum_{p=1}^{N_A} \sum_{l \in D} \frac{\lambda_{pl}}{\lambda} \tau_{pl,jk} P_{Q'}. \quad (4.14)$$

Esta equação pode ser reescrita em função das frequências de despacho, considerando disciplina de prioridades na fila de espera.

$$\bar{T}_{jk} = \frac{\sum_{i=1}^m f_{i,jk}^{[nq]} t_{i,jk}^{[nq]}}{\sum_{i=1}^m f_{i,jk}^{[nq]}} (1 - P_{Q'}) + \frac{\sum_{i=1}^m f_{i,jk}^{[q]} t_{i,jk}^{[q]}}{\sum_{i=1}^m f_{i,jk}^{[q]}} P_{Q'}. \quad (4.15)$$

$$\begin{aligned} \text{Em que } \frac{\sum_{i=1}^m f_{i,jk}^{[q]} t_{i,jk}^{[q]}}{\sum_{i=1}^m f_{i,jk}^{[q]}} P_{Q'} &= \frac{\sum_{i=1}^m \frac{\lambda_{jk}}{\lambda} P_{Q'} \frac{\mu_i}{\mu} \sum_{p=1}^{N_A} \sum_{l \in D} \frac{\lambda_{pl}}{\lambda} \tau_{pl,jk}}{\sum_{i=1}^m \frac{\lambda_{jk}}{\lambda} P_{Q'} \frac{\mu_i}{\mu}} P_{Q'}, \\ &= \frac{\frac{\lambda_{jk}}{\lambda} P_{Q'} \sum_{i=1}^m \frac{\mu_i}{\mu} \sum_{p=1}^{N_A} \sum_{l \in D} \frac{\lambda_{pl}}{\lambda} \tau_{pl,jk}}{\frac{\lambda_{jk}}{\lambda} P_{Q'} \sum_{i=1}^m \frac{\mu_i}{\mu}} P_{Q'}, \text{ como } \sum_{i=1}^m \frac{\mu_i}{\mu} = 1, \text{ tem-se:} \\ &= \sum_{p=1}^{N_A} \sum_{l \in D} \frac{\lambda_{pl}}{\lambda} \tau_{pl,jk} P_{Q'}, \end{aligned}$$

ou seja, a expressão (4.38) equivale à expressão (4.37) em disciplina FCFS.

Tempo médio de viagem ao átomo

O tempo médio de viagem para cada átomo j (\bar{T}_j) é uma medida que, assim como o tempo médio de viagem ao subátomo, reflete o nível de serviço oferecido pelo sistema. Esta equação é dada por:

$$\bar{T}_j = \frac{\sum_{i=1}^m \sum_{k \in D} f_{i,jk}^{[nq]} t_{i,jk}^{[nq]}}{\sum_{i=1}^m \sum_{k \in D} f_{i,jk}^{[nq]}} (1 - P_{Q'}) + \frac{\sum_{i=1}^m \sum_{k \in D} f_{i,jk}^{[q]} t_{i,jk}^{[q]}}{\sum_{i=1}^m \sum_{k \in D} f_{i,jk}^{[q]}} P_{Q'}. \quad (4.16)$$

Tempo médio de viagem para a classe k

Em um sistema com prioridade, outra medida de desempenho importante é o tempo médio de viagem para a classe k (i.e., as classes de prioridades dos chamados). Esta medida pode ser calculada a partir da expressão (4.40).

$$\bar{T}_k = \frac{\sum_{i=1}^m \sum_{j=1}^{N_A} f_{i,jk}^{[nq]} t_{i,jk}^{[nq]}}{\sum_{i=1}^m \sum_{j=1}^{N_A} f_{i,jk}^{[nq]}} (1 - P_{Q'}) + \frac{\sum_{i=1}^m \sum_{j=1}^{N_A} f_{i,jk}^{[q]} t_{i,jk}^{[q]}}{\sum_{i=1}^m \sum_{j=1}^{N_A} f_{i,jk}^{[q]}} P_{Q'}. \quad (4.17)$$

Tempo médio de viagem para cada servidor

O tempo médio de viagem para cada servidor i ($i = 1, \dots, m$) também é uma medida importante para o sistema e reflete o nível de serviço oferecido. Pode ser obtido da seguinte forma:

$$\begin{aligned} \overline{TU}_i &= \frac{\sum_{j=1}^{N_A} \sum_{k \in D} f_{i,jk}^{[nq]} t_{i,jk}^{[nq]} + \sum_{j=1}^{N_A} \sum_{k \in D} f_{i,jk}^{[q]} t_{i,jk}^{[q]}}{\sum_{j=1}^{N_A} \sum_{k \in D} f_{i,jk}^{[nq]} + \sum_{j=1}^{N_A} \sum_{k \in D} f_{i,jk}^{[q]}}, \text{ como:} \\ &= \frac{\sum_{j=1}^{N_A} \sum_{k \in D} f_{i,jk}^{[nq]} t_{i,jk}^{[nq]} + \sum_{j=1}^{N_A} \sum_{k \in D} \frac{\lambda_{jk}}{\lambda} P_{Q'} \frac{\mu_i}{\mu} \sum_{p=1}^{N_A} \sum_{l \in D} \frac{\lambda_{pl}}{\lambda} \tau_{pl,jk}}{\sum_{j=1}^{N_A} \sum_{k \in D} f_{i,jk}^{[nq]} + \sum_{j=1}^{N_A} \sum_{k \in D} \frac{\lambda_{jk}}{\lambda} P_{Q'} \frac{\mu_i}{\mu}} \end{aligned} \quad (4.18)$$

$$= \frac{\sum_{j=1}^{N_A} \sum_{k \in D} f_{i,jk}^{[nq]} t_{i,jk}^{[nq]} + \sum_{j=1}^{N_A} \sum_{k \in D} \sum_{p=1}^{N_A} \sum_{l \in D} \frac{\lambda_{pl} \lambda_{jk}}{\lambda^2} \tau_{pl,jk} P_{Q'} \frac{\mu_i}{\mu}}{\sum_{j=1}^{N_A} \sum_{k \in D} f_{i,jk}^{[nq]} + P_{Q'} \frac{\mu_i}{\mu} \sum_{j=1}^{N_A} \sum_{k \in D} \frac{\lambda_{jk}}{\lambda}}, \text{ tem-se que } \sum_{j=1}^{N_A} \sum_{k \in D} \frac{\lambda_{jk}}{\lambda} = 1.$$

Assim:

$$= \frac{\sum_{j=1}^{N_A} \sum_{k \in D} f_{i,jk}^{[nq]} t_{i,jk}^{[nq]} + (\bar{T}_Q P_{Q'}) \frac{\mu_i}{\mu}}{\sum_{j=1}^{N_A} \sum_{k \in D} f_{i,jk}^{[nq]} + P_{Q'} \frac{\mu_i}{\mu}}.$$

Tempo médio de viagem de cada servidor atendendo usuários da classe k

Em um sistema em que há usuários separados em classes, é interessante calcular o tempo médio de viagem de cada servidor para cada classe k , esta medida pode ser calculada a partir da expressão (4.42).

$$\overline{TU}_{ik} = \frac{\sum_{j=1}^{N_A} f_{i,jk}^{[nq]} t_{i,jk}^{[nq]} + \sum_{j=1}^{N_A} f_{i,jk}^{[q]} t_{i,jk}^{[q]}}{\sum_{j=1}^{N_A} f_{i,jk}^{[nq]} + \sum_{j=1}^{N_A} f_{i,jk}^{[q]}} \quad (4.19)$$

Tempos de Resposta

O tempo de resposta é uma medida importante do ponto de vista do usuário. Este tempo é composto pelo tempo de espera em fila, tempo de *setup* (preparação da equipe) e o tempo de viagem. O tempo de espera em fila pode ser uma medida relevante em sistemas nos quais a taxa de ocupação do sistema é grande o suficiente para a formação de fila de espera, ou seja, $P_Q \gg 0$. O tempo de *setup* é, geralmente, em torno de 1 minuto em sistemas de atendimento emergencial.

Tempos médios de espera para chamados em fila

O tempo de espera para chamados em fila pode ser calculado utilizando a

fórmula de Little ($L_q = \lambda W_q$), que independe da disciplina da fila, apresentada na Seção 3.2. A partir da distribuição de equilíbrio, obtida a partir da resolução do modelo hipercubo, é possível calcular o tempo médio de espera em fila (W_q), calculando o número médio de pessoas em fila (L_q), utilizando o conceito de valor esperado. A Equação (4.43) mostra o tempo médio de espera na fila calculado a partir no número médio de pessoas no sistema (fórmula de Little).

$$\begin{aligned} L_q &= \sum_{s=1}^{n-m} s P_{\{S_s\}} = P_{\{S_4\}} + P_{\{S_5\}} + P_{\{S_6\}}. \\ W_q &= \frac{L_q}{\bar{\lambda}} \end{aligned} \quad (4.20)$$

em que:

$P_{\{S_s\}}$ é a probabilidade de ocorrer o estado S_s da fila de espera, com s usuários em fila;

$$\bar{\lambda} = \lambda (1 - P_{perda});$$

n é o número de usuários no sistema;

m é o número de servidores e;

$$P_{perda} = P_{S_6}.$$

Em um sistema com prioridade, é importante calcular o número médio de usuários em fila e o tempo médio de espera na fila (W_{qk}) para cada classe de usuários (k). Os tempos médios de espera na fila, para cada classe de usuário são calculados utilizando as equações (4.44), (4.45) e (4.46).

$$\begin{aligned} L_{qa} &= 1(P_{\{a\}} + P_{\{ab\}} + P_{\{ac\}} + P_{\{abb\}} + P_{\{abc\}} + P_{\{acc\}}) + 2(P_{\{aa\}} + P_{\{aab\}} + P_{\{aac\}}) + 3(P_{\{aaa\}}), \\ W_{qa} &= \frac{L_{qa}}{\bar{\lambda}_a}. \end{aligned} \quad (4.21)$$

$$\begin{aligned} L_{qb} &= 1(P_{\{b\}} + P_{\{ab\}} + P_{\{bc\}} + P_{\{aab\}} + P_{\{abc\}} + P_{\{bcc\}}) + 2(P_{\{bb\}} + P_{\{abb\}} + P_{\{bbc\}}) + 3(P_{\{bbb\}}), \\ W_{qb} &= \frac{L_{qb}}{\bar{\lambda}_b}. \end{aligned} \quad (4.22)$$

$$L_{qc} = 1(P_{\{c\}} + P_{\{ac\}} + P_{\{bc\}} + P_{\{aac\}} + P_{\{abc\}} + P_{\{bcc\}}) + 2(P_{\{cc\}} + P_{\{acc\}} + P_{\{bcc\}}) + 3(P_{\{ccc\}}),$$

$$W_{qc} = \frac{L_{qc}}{\bar{\lambda}_c} . \quad (4.23)$$

em que:

L_{qa} é o número médio de chamados na fila com prioridade a ;

L_{qb} é o número médio de chamados na fila com prioridade b ;

L_{qc} é o número médio de chamados na fila com prioridade c ;

W_{qa} é o tempo médio de espera na fila para chamados com prioridade a ;

W_{qb} é o tempo médio de espera na fila para chamados com prioridade b ;

W_{qc} é o tempo médio de espera na fila para chamados com prioridade c ;

$\bar{\lambda}_a = \lambda_a (1 - P_{perda})$ é a taxa de entrada dos chamados com prioridade a ;

$\bar{\lambda}_b = \lambda_b (1 - P_{perda})$ é a taxa de entrada dos chamados com prioridade b ;

$\bar{\lambda}_c = \lambda_c (1 - P_{perda})$ é a taxa de entrada dos chamados com prioridade c ;

As expressões (4.44) a (4.46) calculam o tempo médio de espera na fila considerando todos os chamados que entram no sistema, inclusive aqueles com tempo zero de espera em fila. Dessa forma, uma medida interessante é o tempo médio de espera na fila dado que o sistema está saturado:

$$W_q / \{\text{sistema saturado}\} = \frac{W_q}{P_Q} , \quad (4.24)$$

$$W_{qa} / \{\text{sistema saturado}\} = \frac{W_{qa}}{P_Q} , \quad (4.25)$$

$$W_{qb} / \{\text{sistema saturado}\} = \frac{W_{qb}}{P_Q} , \quad (4.26)$$

$$W_{qc} / \{\text{sistema saturado}\} = \frac{W_{qc}}{P_Q} , \quad (4.27)$$

Em que: $P_Q = P_{\{111\}} + P_Q$; e, P_Q é a probabilidade de fila no sistema.

4.3 Resultados computacionais e relação do modelo hipercubo com prioridade com o modelo $M/M/m/C$

Na Seção 3.6, foi ilustrado como o modelo hipercubo clássico se relaciona com o modelo $M/M/m$ a partir dos hiperplanos. Em uma análise similar, o modelo hipercubo com prioridades se relaciona com o modelo $M/M/m/C$ (descrito na Seção

3.5), desde que haja fila finita e os servidores sejam homogêneos. A fim de ilustrar esta relação, é considerado um modelo que representa um sistema com $m = 3$ servidores idênticos, intervalos entre chegadas e de serviço exponencialmente distribuídos e capacidade $C = 6$ usuários, i.e., máximo de 3 usuários em serviço mais 3 usuários na fila, modelo $M/M/m=3/C=6$. As classes de prioridades no modelo hipercubo são representadas pelos subátomos, definidos na Seção 4.1.

Os hiperplanos determinam as probabilidades dos estados do modelo $M/M/m/C$ ($P_{\{1\}} = P_{\{100\}} + P_{\{010\}} + P_{\{001\}}$ e $P_{\{2\}} = P_{\{110\}} + P_{\{011\}} + P_{\{101\}}$). Considere a Figura 4.31, as probabilidades dos estados da fila do modelo $M/M/m/c$ ($S_4 = \{a, b, c\}$, $S_5 = \{aa, ab, ac, bb, bc, cc\}$ e $S_6 = \{aaa, aab, aac, ..., ccc\}$) podem ser calculadas a partir da soma dos estados agregados da cauda do modelo hipercubo, como pode-se ver a seguir:

$$P_{\{S_4\}} = P_{\{a\}} + P_{\{b\}} + P_{\{c\}},$$

$$P_{\{S_5\}} = P_{\{aa\}} + P_{\{ab\}} + P_{\{ac\}} + P_{\{bb\}} + P_{\{bc\}} + P_{\{cc\}} \text{ e,}$$

$$P_{\{S_6\}} = P_{\{aaa\}} + P_{\{aab\}} + P_{\{aac\}} + P_{\{abb\}} + P_{\{abc\}} + P_{\{acc\}} + P_{\{bbb\}} + P_{\{bbc\}} + P_{\{bcc\}} + P_{\{ccc\}}.$$

Para exemplificar numericamente a análise, considere um sistema com $m=3$ servidores idênticos, cada um com taxa de serviço igual a $\mu = 2,1035$ (servidores homogêneos), e as taxas de chegada apresentadas na Tabela 4.3. Os modelos hipercubo e $M/M/3/6$ foram resolvidos de forma independente.

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

Tabela 4.3 – Taxas de chegada (λ) dos átomos (última linha), das classes (última coluna) e dos subátomos (células internas).

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

- ✓ A matriz de preferência de despachos da Tabela 4.1;

- ✓ A matriz dos tempos de viagem $\tau_{il,jk}$ entre subátomos, da Tabela 4.4. Os tempos de viagem desta matriz foram escolhidos arbitrariamente nos átomos 1 e 2, como não há servidores localizados no átomo 3, os tempos de viagem dos subátomos $3k$ a qualquer outro subátomo são iguais a zero;
- ✓ A matriz de localização dos servidores, $l_{m,jk}$, da Tabela 4.5. O servidor 1 representa o VSA, está localizado no átomo 1 e atende aos chamados das classes a , b e c na mesma proporção. Os servidores 2 e 3 representam VSB's, estão localizados no átomo 2 e atendem aos chamados das classes a , b e c na mesma proporção;
- ✓ A matriz $t_{m,jk}$ pode ser obtida multiplicando a matriz l pela matriz τ ,

$$t_{m,jk} = \sum_{j=1}^{N_A} \sum_{k \in C} l_{m,jk} \tau_{jk,il}.$$

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

Tabela 4.4 – Matriz dos tempos de viagem entre os átomos, $\tau_{il,jk}$.

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

Tabela 4.5 – Matriz da localização dos servidores, $l_{m,jk}$.

4.3.1 Medidas de desempenho do modelo hipercubo com prioridade para servidores homogêneos

As probabilidades de estado, obtidas a partir da resolução do modelo hipercubo e os estados da fila, estão na Tabela 4.6. A partir destas probabilidades, foram calculadas as medidas de desempenho para este sistema. Os resultados do *workload*

(carga de trabalho) dos servidores foram calculados a partir da Equação (4.33) e estão na Tabela 4.7.

Estado	$P\{000\}$	$P\{100\}$	$P\{010\}$	$P\{001\}$	$P\{110\}$	$P\{101\}$	$P\{011\}$	$P\{111\}$
Probabilidade	0,1282	0,0358	0,1242	0,0907	0,0539	0,0468	0,1444	0,1598
Estado	$P\{a\}$	$P\{b\}$	$P\{c\}$	$P\{aa\}$	$P\{ab\}$	$P\{ac\}$	$P\{bb\}$	$P\{bc\}$
Probabilidade	0,0013	0,0074	0,0954	0,0000	0,0001	0,0012	0,0004	0,0074
Estado	$P\{cc\}$	$P\{aaa\}$	$P\{aab\}$	$P\{aac\}$	$P\{abb\}$	$P\{abc\}$	$P\{acc\}$	$P\{bbb\}$
Probabilidade	0,0587	0,0000	0,0000	0,0000	0,0000	0,0003	0,0015	0,0000
Estado	$P\{bbc\}$	$P\{bcc\}$	$P\{ccc\}$					
Probabilidade	0,0007	0,0085	0,0332					

Tabela 4.6 – Probabilidade de estado do exemplo ilustrativo.

Ambulâncias	Workload
1	0,5125
2	0,6984
3	0,6579

Tabela 4.7 – *Workloads* das ambulâncias.

Os resultados computacionais das frequências de despacho do servidor i ao subátomo jk , calculados a partir da Equação (4.34), são apresentadas na Tabela 4.8. As frequências de despacho *backup* do sistema foram calculadas a partir das frequências de despacho do servidor i ao subátomo jk . As frequências de despacho *backup* dos servidores (fora de sua área primária) considerando todos os despachos são mostradas na Tabela 4.9. As frequências de despacho *backup* dos subátomos (chamadas atendidas por servidores backup) são mostradas na Tabela 4.10, sendo que a frequência de despacho *backup* total do sistema é 0,5569.

Subátomos										
	Amb	1a	1b	1c	2 ^a	2b	2c	3a	3b	3c
Modelo hipercubo	1	0,0033	0,0147	0,0678	0,0062	0,0084	0,0896	0,0033	0,0074	0,0763
	2	0,0010	0,0143	0,0661	0,0020	0,0127	0,1423	0,0012	0,0112	0,1213
	3	0,0012	0,0269	0,1161	0,0017	0,0093	0,1009	0,0010	0,0082	0,0859

Tabela 4.8 – Frequências de despacho do servidor i para o subátomo jk .

Servidor	Frequências de despacho <i>backup</i>
1	0,9541
2	0,2272
3	0,5928

Tabela 4.9 – Frequências de despacho *backup* do servidor i .

Átomo	1a	1b	1c	2a	2b	2c	3a	3b	3c
Frequências de despacho backup	0,4012	0,5189	0,5355	0,3709	0,5927	0,5831	0,5724	0,5816	0,5722

Tabela 4.10 – Frequências de despacho *backup* do subátomo *jk*.

Tempos de Viagem e de Resposta:

O tempo médio de resposta (discutido na Seção 4.2) é composto do tempo médio de viagem, o tempo médio de *setup* e o tempo médio de espera na fila. As tabelas a seguir apresentam o tempo de resposta sem considerar o tempo de *setup*, que pode ser considerado de 1 a 2 minutos.

A Tabela 4.12 apresenta os resultados dos tempos médios de espera na fila considerando todos os chamados (calculados a partir das equações (4.43), (4.44), (4.45) e (4.46)), e ainda, os tempos médios de espera na fila considerando que o sistema está em fila (calculados a partir das equações (4.47), (4.48), (4.49) e (4.50)). As probabilidades de fila (P_Q) e de saturação (P_{Q^*}) no sistema são de 0,2162 e 0,3760, respectivamente.

Átomo	W_q	W_{qa}	W_{qb}	W_{qc}
Tempo médio de espera na fila (min.)	5,69	3,33	3,51	6,03
Tempo médio de espera na fila / o sistema está em fila (min.)	15,12	8,85	9,34	16,03

Tabela 4.11 – Tempo médio de espera na fila, W_q , W_{qa} , W_{qb} e W_{qc} .

Os tempos médios de viagem (calculados a partir da Equação (4.36)) e resposta do sistema podem ser vistos na Tabela 4.12. A Tabela 4.13 apresenta o número médio de usuários na fila para as classes *a*, *b* e *c*, discutido na Seção 4.2. Além disso, foram calculados os tempos médios de viagem (calculados a partir da Equação (4.37)) e de resposta ao subátomo *jk* considerando todos os chamados, e ainda, os tempos médios de espera na fila considerando que o sistema está em fila.

Tempo médio de viagem (min.)	Tempo médio de resposta (min.)	Tempo médio de resposta / o sistema está em fila (min.)
12,42	18,10	27,54

Tabela 4.12 – Tempo médio de viagem e resposta no sistema.

Átomo	L_q	L_{qa}	L_{qb}	L_{qc}
Número médio de usuários na fila	0,3725	0,0045	0,0260	0,3420

Tabela 4.13 – Tempo médio de espera na fila para os tipos de subátomos a , b e c .

Subátomo	1a	1b	1c	2a	2b	2c	3a	3b	3c
Tempo médio de viagem (min.)	11,20	11,50	11,50	12,18	11,70	11,70	14,40	14,16	14,16
Tempo médio de resposta (min.)	14,52	15,01	17,52	15,51	15,21	17,72	17,72	17,67	20,18
Tempo médio de resposta/ o sistema está em fila (min.)	23,37	24,35	33,55	24,35	24,55	33,75	26,57	27,00	36,21

Tabela 4.14 – Tempo médio de viagem e de resposta ao subátomo jk .

Os tempos médios de viagem, calculados a partir da Equação (4.37), e de resposta considerando todos os chamados, e ainda, os tempos médios de espera na fila considerando que o sistema está em fila estão na Tabela 4.15. Os tempos médios de viagem, calculados a partir da Equação (4.41), e de resposta do servidor i , considerando todos os chamados, e ainda, os tempos médios de espera na fila considerando que o sistema está em fila estão na Tabela 4.16. Também foram calculados os tempos médios de viagem, calculados a partir da Equação (4.42), e de resposta do servidor i para cada classe k considerando todos os chamados, e ainda, os tempos médios de espera na fila considerando que o sistema está em fila. Os resultados estão na Tabela 4.17.

Subátomo	a	b	c
Tempo médio de viagem (min.)	12,16	11,78	12,17
Tempo médio de resposta (min.)	15,49	15,29	18,19
Tempo médio de resposta/ o sistema está em fila (min.)	24,34	24,63	34,22

Tabela 4.15 – Tempo médio de viagem e de resposta aos tipos de subátomos a , b e c .

Ambulâncias	1	2	3
Tempo de Viagem (min.)	12,69	12,38	12,24
Tempo de Resposta (min.)	18,37	18,07	17,93
Tempo médio de resposta/ o sistema está em fila(min.)	27,22	27,41	33,95

Tabela 4.16 – Tempo médio de viagem e de resposta do servidor i .

Átomo	1a	1b	1c	2a	2b	2c	3a	3b	3c
Tempo de Viagem (min.)	12,26	12,16	12,16	11,76	12,06	11,78	12,40	12,21	12,17
Tempo médio de resposta (min.)	15,59	15,67	18,19	15,09	15,57	17,81	15,73	15,72	18,19
Tempo médio de resposta/ o sistema está em fila (min.)	24,43	25,01	34,22	23,94	24,91	33,83	24,57	25,06	34,22

Tabela 4.17 – Tempo de viagem e de resposta do servidor i para os tipos de subátomos a , b e c .

4.3.2 Medidas de desempenho do modelo $M/M/m/C$ e comparação com o modelo hipercubo com prioridade do exemplo ilustrativo

Conforme mencionado anteriormente, o modelo hipercubo com prioridade na fila e classes de usuários relaciona-se com o modelo $M/M/m/C$ a partir da análise dos hiperplanos do hipercubo e dos estados agregados da fila de espera. Para ilustrar isso, foram comparados os resultados do modelo hipercubo utilizando servidores homogêneos, mostrado na Seção 4.3.1 com os resultados do modelo $M/M/m/C$ com parâmetros: $m = 3$; $C = 6$; taxa de chegada total dos chamados obtida a partir da Tabela 4.3 ($\lambda = 4,1129$); taxa de serviço $\mu = 2,1035$, assim como no modelo hipercubo.

Os resultados das probabilidades de estado dos dois modelos estão na Tabela 4.18. Na primeira coluna estão os resultados do modelo $M/M/m=3/C=6$, por exemplo: $P_{\{1\}} = 0,2507$ é a probabilidade de haver 1 usuário no sistema (1 servidor ocupado) e $P_{\{4\}} = 0,1041$ é a probabilidade de haver 4 usuários no sistema. A segunda coluna mostra os resultados das probabilidades de estado obtidas a partir do modelo hipercubo com prioridade da Seção 4.3.1 e as probabilidades de estado agregadas. Pode-se notar que as probabilidades dos estados da fila $M/M/m/C$ podem ser obtidos a partir dos estados agregados do modelo hipercubo com prioridade. Por exemplo, a soma

probabilidades dos servidores 1, 2 e 3 estarem ocupados ($P_{\{100\}}$, $P_{\{010\}}$, $P_{\{001\}}$) do modelo hipercubo com prioridade, é igual a probabilidade de um servidor estar ocupado em um sistema $M/M/m/C$ ($P_{\{100\}} + P_{\{010\}} + P_{\{001\}} = 0,0358 + 0,1242 + 0,0907 = 0,2507 = P_{\{1\}}$). O mesmo acontece para os estados agregados da fila, a soma das probabilidades do modelo hipercubo com prioridade de haver um usuário na fila com prioridade a , b ou c é igual a probabilidade de haver um usuário na fila em um sistema $M/M/m/C$ ($P_{\{a\}} + P_{\{b\}} + P_{\{c\}} = 0,0013 + 0,0074 + 0,0954 = 0,1041 = P_{\{S_4\}}$).

As medidas de desempenho dos dois sistemas foram calculadas e os resultados estão na Tabela 4.19. Os resultados das probabilidade dos estados agregados do modelo hipercubo e as probabilidades dos estados do modelo $M/M/m/C$ são os mesmos.

Modelo $M/M/m/C$	Modelo hipercubo com prioridade
$P_{\{0\}} = 0,1282$	$P_{\{000\}} = \mathbf{0,1282}$
$P_{\{1\}} = 0,2507$	$P_{\{100\}} + P_{\{010\}} + P_{\{001\}} =$ $0,0358 + 0,1242 + 0,0907 = \mathbf{0,2507}$
$P_{\{2\}} = 0,2451$	$P_{\{110\}} + P_{\{101\}} + P_{\{011\}} =$ $0,0539 + 0,0468 + 0,1444 = \mathbf{0,2451}$
$P_{\{3\}} = 0,1598$	$P_{\{111\}} = \mathbf{0,1598}$
$P_{\{4\}} = 0,1041$	$P_{\{S_4\}} = P_{\{a\}} + P_{\{b\}} + P_{\{c\}} =$ $0,0013 + 0,0074 + 0,0954 = \mathbf{0,1041}$
$P_{\{5\}} = 0,0679$	$P_{\{S_5\}} = P_{\{aa\}} + P_{\{ab\}} + P_{\{ac\}} + P_{\{bb\}} + P_{\{bc\}} + P_{\{cc\}} =$ $0,0000 + 0,0001 + 0,0012 + 0,0004 + 0,0074 + 0,0587 = \mathbf{0,0679}$
$P_{\{6\}} = 0,0442$	$P_{\{S_6\}} = P_{\{aaa\}} + P_{\{aab\}} + P_{\{aac\}} + P_{\{abb\}} + P_{\{abc\}} + P_{\{acc\}} + P_{\{bbb\}} + P_{\{bbc\}} + P_{\{bcc\}} + P_{\{ccc\}} =$ $0,0000 + 0,0000 + 0,0000 + 0,0000 + 0,0003 + 0,0015 + 0,0000 + 0,0007 + 0,0085 + 0,0332 = \mathbf{0,0442}$
$P_q = 0,2162$	$P_q = P_{\{S_4\}} + P_{\{S_5\}} + P_{\{S_6\}} =$ $0,1041 + 0,0679 + 0,0442 = \mathbf{0,2162}$

Tabela 4.18 – Relação das probabilidades dos estados do modelo $M/M/m/C$ com os estados do modelo hipercubo.)

Medidas de desempenho	$M/M/m/C$	Modelo hipercubo
L_q	0,37	0,37
L	2,24	2,24
W_q	5,69 min.	5,69 min.
W	34,21 min.	34,21 min.

Tabela 4.19 – Comparação das medidas de desempenho do modelo $M/M/m/C$ com o modelo hipercubo.

4.3.3 Medidas de desempenho do modelo $M_r/M/m$ e do modelo hipercubo com prioridade do exemplo ilustrativo

Como descrito na Seção 2.4.1, o modelo $M_r/M/m$ considera m servidores homogêneos e fila infinita. Em um sistema com taxas de chegada $\lambda_a = 0,0851$, $\lambda_b = 0,4651$ e $\lambda_c = 3,5627$ (obtidos a partir da Tabela 4.3) e taxa de serviço $\mu = 2,1035$, o tempo médio de espera na fila pode ser calculado a partir da Equação (2.46). Os resultados dos tempos médios de espera para este modelo estão na segunda coluna da Tabela 4.20. A terceira coluna desta tabela mostra os resultados do tempo médio de espera na fila (para o sistema), calculados a partir do modelo hipercubo com prioridade da Seção 4.3.1.

Medidas de desempenho	$M_r/M/m$	Modelo hipercubo	Simulação hipercubo	Desvio (%)
W_{qa}	4,0830	3,3269	3,4026	-2,2754
W_{qb}	4,4730	3,5111	3,5502	-1,1136
W_{qc}	12,6711	6,0264	6,0342	-0,1294

Tabela 4.20 – Comparação do modelo hipercubo com o modelo $M_r/M/m$.

Como o modelo $M_r/M/m$ considera fila infinita e o modelo hipercubo com prioridade considera fila finita, a comparação do tempo médio de espera na fila dos dois modelos fica comprometida. Porém, espera-se que em um sistema de fila infinita o tempo de espera na fila seja maior que um sistema com fila finita, pois este último trabalha com perda de usuários.

Um modelo de simulação foi desenvolvido e implementado no *software* Arena para representar o mesmo problema descrito pelo modelo hipercubo (os detalhes da simulação estão no Anexo D). Os resultados dos tempos médios de espera na fila obtidos com simulação estão na quarta coluna da Tabela 4.20. A quinta coluna apresenta os desvios do modelo hipercubo com a simulação. Espera-se que os resultados da simulação e do modelo teórico sejam convergentes a menos de um erro amostral. Pode-se verificar na tabela que todos os desvios dos modelos e da simulação foram relativamente pequenos, menores que 3%.

4.3.4 Modelo hipercubo com prioridade com servidores heterogêneos e comparação com o modelo de simulação do exemplo ilustrativo

O exemplo ilustrativo foi resolvido numericamente, considerando taxa média de ocupação igual a 0,6152. Esta taxa foi escolhida a fim de representar a taxa observada no SAMU-RP original. As matrizes $\tau_{pl,jk}$, $l_{i,jk}$ e $t_{i,jk}$ utilizadas podem ser vistas nas Tabelas 4.3, 4.4 e 4.5, respectivamente. As taxas de chegada do usuário e as taxas de serviço no sistema são mostradas nas Tabelas 4.3 e 4.21, respectivamente.

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

Tabela 4.21 – Taxa de serviço para os três servidores.

Além disso, o mesmo modelo de simulação utilizado na seção anterior foi aplicado, porém utilizando as taxas de serviço da Tabela 4.21 com o objetivo de validar o modelo analítico. Os resultados dos dois modelos devem convergir, a menos de um erro amostral. O modelo hipercubo foi resolvido pelo método exato e foram obtidas as probabilidades de estado do sistema. Os resultados podem ser vistos na Tabela 4.22.

Estado	$P\{000\}$	$P\{100\}$	$P\{010\}$	$P\{001\}$	$P\{110\}$	$P\{101\}$	$P\{011\}$	$P\{111\}$
Probabilidade	0,1352	0,0550	0,1210	0,0780	0,0710	0,0536	0,1164	0,1571
Estado	$P\{S_a^1\}$	$P\{S_b^1\}$	$P\{S_c^1\}$	$P\{S_{a,a}^2\}$	$P\{S_{a,b}^2\}$	$P\{S_{a,c}^2\}$	$P\{S_{b,b}^2\}$	$P\{S_{b,c}^2\}$
Probabilidade	0,0013	0,0073	0,0938	0,0000	0,0001	0,0012	0,0003	0,0073
Estado	$P\{S_{c,c}^2\}$	$P\{S_{a,a,a}^3\}$	$P\{S_{a,a,b}^3\}$	$P\{S_{a,a,c}^3\}$	$P\{S_{a,b,b}^3\}$	$P\{S_{a,b,c}^3\}$	$P\{S_{a,c,c}^3\}$	$P\{S_{b,b,b}^3\}$
Probabilidade	0,0578	0,0000	0,0000	0,0000	0,0000	0,0003	0,0015	0,0000
Estado	$P\{S_{b,b,c}^3\}$	$P\{S_{b,c,c}^3\}$	$P\{S_{c,c,c}^3\}$					
Probabilidade	0,0007	0,0084	0,0326					

Tabela 4.22 – Probabilidade de estado.

A partir da probabilidade de estado dos servidores do sistema, foram calculadas as medidas de desempenho de interesse. Na Tabela 4.23, é apresentada a comparação da carga de trabalho (*workload*) dos servidores (calculados a partir da Equação (4.33)) com a simulação. Os desvios foram relativamente pequenos, menores que 0,4%.

Os resultados computacionais das frequências de despacho do servidor i ao

subátomo jk (calculados a partir da Equação (4.34)), para o caso dos servidores homogêneos e heterogêneos, são mostrados na Tabela 4.35.

Ambulância	Servidores heterogêneos	Simulação	Desvio (%)
1	0,5494	0,5489	0,0911
2	0,6782	0,6759	0,3403
3	0,6179	0,6162	0,2759

Tabela 4.23 – *Workloads* das ambulâncias.

Subátomos										
	Ambulância	1a	1b	1c	2a	2b	2c	3a	3b	3c
Servidores heterogêneos	1	0,0028	0,0113	0,0514	0,0055	0,0064	0,0681	0,0028	0,0055	0,0579
	2	0,0011	0,0143	0,0669	0,0023	0,0135	0,1519	0,0014	0,0119	0,1295
	3	0,0016	0,0303	0,1316	0,0020	0,0105	0,1128	0,0013	0,0093	0,0961

Tabela 4.24 – Frequências de despacho do servidor i para o subátomo j .

As frequências de despacho *backup* do sistema foram calculadas a partir das frequências de despacho do servidor i ao átomo j . A frequência de despacho *backup* total do sistema na simulação é de 0,5224, no modelo é 0,5149, com desvio de -1,46%. As frequências de despacho *backup* dos servidores (fora de sua área primaria) considerando todos os despachos, para o modelo hipercubo com servidores heterogêneos, são mostradas na Tabela 4.25. As frequências de despacho *backup* dos subátomos (chamadas atendidas por servidores *backup*), para o caso dos servidores heterogêneos são mostradas na Tabela 4.26.

	Servidores heterogêneos
$F_{J1} =$	0,9473
$F_{J2} =$	0,2188
$F_{J3} =$	0,5903

Tabela 4.25 – Frequências de despacho *backup* do servidor i .

Subátomo	Servidores heterogêneos
1a	0,4760
1b	0,4573
1c	0,4732
2a	0,4439
2b	0,5540
2c	0,5434
3a	0,4770
3b	0,5525
3c	0,5433

Tabela 4.26 – Frequências de despacho *backup* do subátomo jk .

Tempos de Viagem:

A Tabela 4.27 mostra os resultados dos tempos médios de espera na fila para todos os chamados do sistema (W_q) para as classes a , b e c (W_{qa} , W_{qb} e W_{qc} calculados a partir das equações (4.44), (4.45) e (4.46)), para o caso dos servidores heterogêneos e a simulação. Os desvios foram pequenos em todas as situações, o maior desvio encontrado foi de 2,36%.

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

Tabela 4.27 – Tempo médio de espera na fila W_q , W_{qa} , W_{qb} e W_{qc} .

O tempo médio de viagem (calculados a partir da Equação (4.36)) e de resposta no sistema, juntamente com a comparação com a simulação, são apresentados na Tabela 4.28. O tempo de *setup* foi considerado constante (1 minuto, como no exemplo ilustrativo 1). Os desvios do modelo em relação a amostra são pequenos, o maior deles é 3,12%.

	Tempo de Viagem (min.)	Tempo de Resposta para chamados em fila (min.)
Servidores heterogêneos	12,46	18,05
Simulação	12,08	17,61
Desvio (minutos)	0,38	0,44
Desvio (%)	3,12	2,49

Tabela 4.28 – Tempo médio de viagem e resposta no sistema.

Foram calculados os tempos médios de viagem (calculados a partir da Equação 4.38) e de resposta ao átomo jk , para o caso dos servidores heterogêneos e comparação com a simulação, como se pode ver na Tabela 4.29. Os desvios são relativamente pequenos, todos menores que 4%.

Subátomo	Tempo de Viagem				Tempo de Resposta para chamados em fila			
	Servidores heterogêneos (min.)	Simulação (min.)	Desvio (min.)	Desvio (%)	Servidores heterogêneos (min.)	Simulação (min.)	Desvio (min.)	Desvio (%)
1a	11,02	10,55	0,47	4,43	14,29	13,75	0,54	3,95
1b	11,25	10,78	0,47	4,39	14,71	14,19	0,52	3,65
1c	11,25	10,78	0,47	4,39	17,18	16,64	0,54	3,24
2a	11,86	11,53	0,33	2,86	15,13	14,72	0,40	2,75
2b	11,43	11,08	0,35	3,12	14,88	14,49	0,39	2,69
2c	11,43	11,08	0,35	3,16	17,35	16,94	0,42	2,45
3a	15,07	14,77	0,30	2,04	18,34	17,96	0,38	2,10
3b	14,85	14,53	0,33	2,26	18,31	17,93	0,37	2,08
3c	14,85	14,54	0,32	2,17	20,78	20,40	0,38	1,87

Tabela 4.29 – Tempo médio de viagem e de resposta ao subátomo jk .

Os tempos médios de viagem (calculados a partir da Equação (4.40)) e de resposta para as classes a , b e c estão na Tabela 4.30. Os tempos médios de viagem (calculados a partir da Equação (4.41)) e de resposta do servidor i , estão na Tabela 4.42. Na Tabela 4.43 estão os resultados dos tempos médios de viagem (calculados a partir da Equação (4.42)) e de resposta do servidor i para as classes a , b e c . Em todos os casos, foi feita a comparação do modelo hipercubo com a simulação. Os desvios são relativamente pequenos, todos menores que 4%.

Classes	Tempo de Viagem				Tempo de Resposta para chamados em fila			
	Servidores heterogêneos (min.)	Simulação (min.)	Desvio (min.)	Desvio (%)	Servidores heterogêneos (min.)	Simulação (min.)	Desvio (min.)	Desvio (%)
a	12,50	12,11	0,38	3,17	15,77	15,31	0,46	3,00
b	12,14	11,67	0,46	3,98	15,59	15,08	0,51	3,38
c	12,50	12,14	0,36	3,00	18,43	18,00	0,43	2,39

Tabela 4.30 – Tempo médio de viagem e de resposta para as classes a , b e c .

Ambulância	Tempo de Viagem				Tempo de Resposta para chamados em fila			
	Servidores heterogêneos (min.)	Simulação (min.)	Desvio (min.)	Desvio (%)	Servidores heterogêneos (min.)	Simulação (min.)	Desvio (min.)	Desvio (%)
1	12,72	12,32	0,39	3,20	18,31	17,85	0,46	2,55
2	12,45	12,08	0,37	3,08	18,04	17,60	0,43	2,46
3	12,34	11,97	0,37	3,09	17,92	17,49	0,43	2,46

Tabela 4.31 – Tempo médio de viagem e de resposta do servidor i .

	Tempo de Viagem				Tempo de Resposta para chamados em fila			
Ambulância - Classe	Servidores heterogêneos (min.)	Simulação (min.)	Desvio (min.)	Desvio (%)	Servidores heterogêneos (min.)	Simulação (min.)	Desvio (min.)	Desvio (%)
1 - a	12,71	12,29	0,42	3,38	15,98	15,49	0,49	3,17
1 - b	12,51	11,55	0,96	8,35	15,96	14,95	1,01	6,75
1 - c	12,40	12,43	-0,02	-0,19	18,32	18,28	0,04	0,23
2 - a	11,72	11,71	0,01	0,11	14,99	14,90	0,09	0,59
2 - b	12,60	12,36	0,24	1,91	16,05	15,77	0,28	1,78
2 - c	12,24	12,05	0,19	1,60	18,17	17,91	0,26	1,44
3 - a	12,42	12,08	0,34	2,80	15,69	15,28	0,41	2,71
3 - b	12,43	11,18	1,25	11,21	15,88	14,59	1,30	8,90
3 - c	12,80	12,08	0,72	5,96	18,72	17,94	0,79	4,38

Tabela 4.32 – Tempo de viagem e de resposta do servidor i para as classes a , b e c .

Caso o tamanho da fila seja de no máximo 5 usuários, os desvios são ainda menores, por exemplo, as Tabelas 4.33 e 4.34 mostram os *workloads*, com desvios menores que 0,3% e os tempos médios de espera na fila, com desvios menores que 1,3% para o modelo e a simulação, respectivamente.

Ambulância	Servidores heterogêneos	Simulação	Desvio (%)
1	0,57	0,57	0,00
2	0,69	0,69	0,21
3	0,64	0,63	0,15

Tabela 4.33 – *Workload* dos servidores para até 5 usuários na fila.

	Servidores heterogêneos (min.)	Simulação (min.)	Desvio (min.)	Desvio (%)
W_q	8,12	8,11	0,00	0,04
W_{qa}	3,72	3,76	-0,03	-0,89
W_{qb}	4,02	3,97	0,05	1,22
W_{qc}	8,76	8,76	0,00	0,01

Tabela 4.34 – Tempo médio de espera na fila W_q , W_{qa} , W_{qb} e W_{qc} , para até 5 usuários na fila.

Convém ressaltar que a principal diferença do modelo hipercubo considerando prioridade na fila e o modelo hipercubo clássico está na consideração de prioridade de atendimento nos estados da cauda (fila) do hipercubo, o que resulta em diferenças na avaliação dos tempos de resposta em medidas de desempenho das classes de usuários. Isso é particularmente importante em sistemas em que a probabilidade de fila não é muito pequena. Por exemplo, os tempos de espera na fila e os tempos de

resposta que consideram as classes de usuários são relativamente diferentes nas duas abordagens. Para ilustrar isso, na Tabela 4.35 estão apresentados os tempos médios de espera na fila, calculados pelo modelo hipercubo clássico e pelo modelo hipercubo com prioridade na fila. No modelo hipercubo clássico, apenas o tempo de espera na fila de todos os chamados é calculado, enquanto que no modelo hipercubo, considerando prioridade na fila, também podem ser calculados os tempos de espera para cada classe de usuário.

	Modelo com prioridade	Modelo Clássico	Desvio (minutos)	Desvio (%)
W_q	5,59	5,59	0,00	0,00
W_{qa}	3,27	5,59	-2,32	-41,49
W_{qb}	3,45	5,59	-2,14	-38,25
W_{qc}	5,92	5,59	0,33	5,98

Tabela 4.35 – Tempo de espera na fila do modelo hipercubo clássico, considerando prioridade na fila

A Tabela 4.36 mostra os tempos de viagem e de resposta para cada subátomo do sistema. Os tempos de viagem calculados por ambos os modelos são bastante próximos. Já os tempos de resposta, são bastante diferentes. No modelo hipercubo clássico, o tempo de resposta é aproximado pela soma dos tempos de viagem, mais o tempo médio de espera na fila de todos os chamados, ao passo que no modelo hipercubo, considerando prioridade na fila, os tempos de resposta são obtidos pelos tempos de viagem mais o tempo médio de espera na fila considerando cada classe de usuário. Pode-se observar que, principalmente nas classes *a*, a diferença é significativa (da ordem de 2 a 3 minutos), uma vez que estes usuários se encontram em risco de vida.

	Tempo de Resposta para chamados em fila (min.)			
Subátomo	Modelo com prioridade	Modelo Clássico	Desvio (minutos)	Desvio (%)
1a	14,29	16,61	-2,32	-13,96
1b	14,71	16,84	-2,14	-12,69
1c	17,18	16,84	0,33	1,99
2a	15,13	17,44	-2,32	-13,29
2b	14,88	17,02	-2,14	-12,56
2c	17,35	17,02	0,33	1,97
3a	18,34	20,66	-2,32	-11,22
3b	18,31	20,44	-2,14	-10,46
3c	20,78	20,44	0,33	1,64

Tabela 4.36 – Tempo de viagem e de resposta do modelo hipercubo clássico, considerando prioridade na fila

Essa seção dedicou-se à comparação do modelo hipercubo com os modelos $M/M/m/C$ e $M_r/M/m$, assim como o modelo hipercubo com um modelo de simulação para o caso de servidores heterogêneos. Os resultados mostraram que, caso o sistema tenha servidores homogêneos e fila finita, a distribuição de equilíbrio da fila e as medidas de desempenho agregadas (como L_q , L , W_q e W) são as mesmas, tanto no modelo hipercubo como no modelo $M/M/m/C$. A partir do modelo $M_r/M/m$ é possível calcular os tempos médios de espera na fila apenas em sistemas com servidores homogêneos e fila infinita, de forma que a comparação com o modelo hipercubo (que considera fila finita) é comprometida nos casos em que a probabilidade de perda é maior que zero. Pelo fato de o modelo hipercubo, considerando disciplina de prioridade na fila, poder considerar sistemas de servidores homogêneos ou heterogêneos, pode ser considerado um modelo mais geral do que os modelos $M/M/m/C$ e $M_r/M/m$. Os desvios pequenos (em geral, menores que 5%) dos modelos em relação à simulação mostram que os resultados obtidos são robustos.

4.4 Generalização das equações da fila do modelo hipercubo considerando prioridade na fila

Devido à dificuldade para se apresentar, de forma analítica, a generalização das equações de equilíbrio (4.32) do final da Seção 4.1, para o caso com r classes de prioridades de chamados (ao invés de apenas $r = 3$ como na Seção 4.2) e com tamanho de fila limitado em $s = n - m$, em que n é o número de usuários no sistema e m é o número de servidores (ao invés de apenas $s = 3$, como na Seção 4.2), nesta seção, apresenta-se um procedimento (em pseudocódigo) que generaliza a geração destas equações para r classes de prioridades e capacidade de fila s .

Para representar os estados da fila com prioridades, são utilizados vetores compostos por letras. Por exemplo, o estado $\{a\}$ representa um usuário na fila com mais alta prioridade, o estado $\{ab\}$ representa dois usuários na fila, sendo que o usuário a tem prioridade no atendimento sobre o usuário b , e assim por diante. Dessa forma, os estados da fila são compostos por vetores de *strings* e colocados em ordem alfabética, de acordo com a quantidade de usuários na fila e de classes.

Para exemplificar o procedimento, é considerado um exemplo com apenas três classes de usuários. Os estados da fila com uma chamada são: a , b e c . Os estados da fila com dois usuários são obtidos da seguinte forma: acrescenta-se " a " à esquerda de

todos os estados da fila com uma chamada; acrescenta-se "b" à esquerda dos estados da fila com uma chamada que começam com "b" ou "c" e acrescenta-se "c" à esquerda dos estados da fila com uma chamada que começam com "c". Esse procedimento é repetido para estados da fila com mais chamados na fila. O pseudocódigo que gera os estados da fila é apresentado a seguir, cada s usuário na fila gera um nível de possíveis estados. Os termos que se encontram entre “//” são comentários. Sejam:

r o número de classes de prioridade
 s o tamanho máximo da fila
 $c(i)$ o símbolo da i -ésima classe de prioridade em ordem lexicográfica;
 $n(j)$ o número de estados em fila de j chamadas;
 $e(i,j)$ o j -ésimo estado da fila com i chamadas.

Procedimento para gerar os estados de fila:

```
{
for  $i = 1$  to  $r$  do
     $e(1,i) = c(i)$  //Os estados no primeiro nível da fila são as classes de usuários representadas por
                    letras  $\{c(i)\}$ //
for  $i=2$  to  $s$  do //O procedimento a seguir é repetido para cada nível da fila//
     $m = 0$ 
    for  $j = 1$  to  $r$  do;
        for  $k = 1$  to  $n(i-1)$  do //Os estados do nível  $i$  são gerados a partir dos estados do nível
                                imediatamente anterior, colocados em ordem lexicográfica//
            if  $c(j) \leq e(i-1,k)$  // O procedimento a seguir é realizado somente se no estado do nível
                                imediatamente anterior houver apenas usuários de prioridade maior ou
                                igual a  $c(j)$ //
                 $m = m+1$  //Posição de um estado no nível  $i$ //
                 $e(i,m) = c(j) + e(i-1,j)$  //Um usuário da classe  $j$  é colocado a esquerda do estado do nível
                                        anterior  $(i-1)$ //
    }
}
```

O pseudocódigo descrito anteriormente gera todos os estados de uma fila com r classes de usuários e capacidade s da fila. A Figura 4.31 ilustra o procedimento descrito anteriormente para uma fila com $r = 3$ classes de usuários e limite de comprimento de cauda $s = 3$

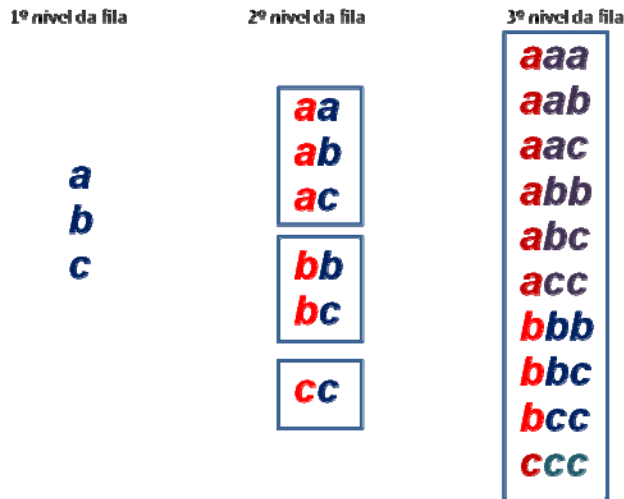


Figura 4.8 – Exemplo de geração dos estados da fila para $r = 3$.

Assim, temos que:

$$B_p = \left(\{a\}, \{b\}, \{c\}, \{aa\}, \{ab\}, \{ac\}, \{bb\}, \{bc\}, \{cc\}, \{aaa\}, \{aab\}, \{aac\}, \{abb\}, \{abc\}, \{acc\}, \{bbb\}, \{bbc\}, \{bcc\}, \{ccc\} \right).$$

Como descrito na Seção 3.3, a taxa com que o sistema sai do estado n (lado esquerdo das equações) é igual à taxa com que o sistema entra no estado n (lado direito das equações). A lógica encontrada para gerar os coeficientes dos estados da fila, em que as entradas são representadas por “+” e as saídas por “-” é apresentada a seguir:

1. Comparar estados dois a dois;
2. Eliminar pares com diferenças, entre o número de clientes, diferentes de +1 ou -1;
3. Retirar do estado com mais clientes todos os clientes do estado com menos clientes e ver o que sobrou;
4. Se sobrou mais de um cliente, deve-se eliminar o par;
5. Se o estado de origem tem menos clientes, o par corresponde à chegada do cliente que sobrou, como, por exemplo, de ‘ab’ para ‘aab’ com entrada de ‘a’;
6. Se o estado de origem tem mais clientes, há um filtro adicional para eliminar transições inviáveis, como, por exemplo, de ‘ab’ para ‘a’;
7. O par que passa no filtro representa as transições como, por exemplo, de ‘bcc’ para ‘cc’, com saída da fila de ‘b’.

O pseudocódigo a seguir gera a taxa com que o sistema entra em um estado n . Sejam:

$n(k)$ o número de estados na última camada da fila;

nt o número total de estados na fila até a última chamada.

$S1$ e $S2$ os vetores (string) de possíveis estados da fila

$L1$ e $L2$ os respectivos tamanhos dos vetores $S1$ e $S2$ (quantidade de usuários na fila)

$InOut$ guarda a informação se entrou ou saiu usuários na fila com os sinais $+$ e $-$, respectivamente.

$Bigger$ o estado com mais usuários.

$Smaller$ o estado com menos usuários.

$User$ um vetor obtido da diferença entre o estado maior e o menor (da comparação entre $S1$ e $S2$), o que sobra é o usuário que entrou ou saiu do sistema (esta diferença deve ser sempre de 1 usuário).

Procedimento para gerar taxa com que sistema sai de um estado n :

```
{
for  $i = 1$  to  $nt$  do
    for  $j = 1$  to  $nt$  do        //Procedimento para gerar uma matriz quadrada (matriz de string) com o
                                número de linhas e colunas igual ao número total de estados na fila até a
                                última chamada //
    {
         $L1 = Length(S(i))$       //Calcula a quantidade de usuários na fila do estado  $\{S(i)\}$  //
         $L2 = Length(S(j))$       //Calcula a quantidade de usuários na fila do estado  $\{S(j)\}$  //
        if  $|L1-L2| < 1$           //O procedimento a seguir é realizado somente se a diferença (em módulo)
                                do número de usuários na fila dos dois estados forem diferentes de 1//

         $mat(i,j) = empty$ 
        GoTo End

        if  $L1 > L2$               //Se o estado  $S(i)$  tiver um usuário a mais que  $S(j)$ , então //
             $InOut = +$            // O sinal de mais indica que alguém entrou na fila //
             $Bigger = S(i)$        //O estado com mais usuários é o  $S(i)$  //
             $Smaller = S(j)$      //O estado com menos usuários é o  $S(j)$  //
        Else                    //Se o estado  $S(i)$  tiver um usuário a menos que  $S(j)$ , então //
             $InOut = -$            // O sinal de menos indica que alguém saiu da fila //
             $Bigger = S(j)$        //O estado com mais usuários é o  $S(j)$  //
             $Smaller = S(i)$      //O estado com menos usuários é o  $S(i)$  //

         $user = Bigger - Smaller$  //Este procedimento (deve ser construída uma rotina que faça esta
                                operação) indica uma diferença de strings, no sentido de que, todos os
```

```

usuários que estão ao mesmo tempo nos dois estados são retirados do
maior estado na fila//
if Length(user) <> 1 //Se da diferença anterior sobrar mais que um usuário, então //
    mat(i,j) = empty
    GoTo End
if (InOut = -) and (user > Bigger) //Se um usuário saiu da fila (sinal -) e o usuário que sobrou
da diferença de strings for maior que o maior estado da fila
(esteste procedimento garante que um usuário com prioridade
menor nunca será atendido se um usuário com maior
prioridade estiver na fila), então//
    mat(i,j) = empty
    GoTo End
mat(i,j) = InOut+user //O elemento da matriz na linha i e coluna j é o sinal (mais ou menos,
indicando se entrou ou saiu alguém da fila, respectivamente) mais o
usuário que sobrou da diferença de strings//
End
}

```

O sistema sai de um estado n quando há um chamado, com taxa λ (para estados que não representam a capacidade máxima da fila), ou quando há um término de serviço, com taxa μ . A taxa em que o sistema sai de um estado n é colocada na diagonal da matriz dos coeficientes das equações de equilíbrio. O procedimento a seguir é feito para a diagonal da matriz.

Procedimento para a diagonal da matriz

```

{
j = nt - n(k) //Quantidade de estados que não estão no último nível da fila //
for i = 1 to nt do //O procedimento é repetido para todos os estados da fila //
    if i <= j //Se os estados da fila não são do último nível da fila, então //
        mat(i,i) = totLamb+mu
    else
        mat(i,i) = mu
}

```

Como descrito na Seção 3.5, o sistema gerado pelas equações de equilíbrio da fila pode ser resolvido independentemente das equações do modelo hipercubo, com a informação da probabilidade de todos os servidores estarem ocupados. Os termos independentes das equações são construídos a partir do seguinte pseudocódigo:

Procedimento para completar a última coluna da matriz (termos independentes das equações de equilíbrio):

```
{
for i = 1 to nt do
  if i <= r      //Se os estados forem do primeiro nível da fila, então//
    mat(i,nt+1) = -Lambda(i)*P{11...1}
  else
    mat(i,nt+1) = 0.0
}
```

Os procedimentos foram codificados e implementados computacionalmente em Pascal, reproduzindo adequadamente o sistema de equações de equilíbrio do sistema em diversos testes. A Figura 4.32 mostra a matriz obtida a partir dos procedimentos descritos anteriormente.

$$\begin{array}{c}
 \begin{bmatrix} a & b & c & aa & ab & ac & bb & \dots & bcc & ccc \end{bmatrix} \\
 \begin{bmatrix} a \\ b \\ c \\ aa \\ ab \\ ac \\ bb \\ \vdots \\ bcc \\ ccc \end{bmatrix}
 \end{array}
 \begin{bmatrix}
 \lambda + \mu & \square & \square & -a & \square & \square & \square & \dots & \square & \square & \lambda_a P_{\{111\}} \\
 \square & \lambda + \mu & \square & \square & -a & \square & -b & \dots & \square & \square & \lambda_b P_{\{111\}} \\
 \square & \square & \lambda + \mu & \square & \square & -a & \square & \dots & \square & \square & \lambda_c P_{\{111\}} \\
 +a & \square & \square & \lambda + \mu & \square & \square & \square & \dots & \square & \square & \square \\
 +b & +a & \square & \square & \lambda + \mu & \square & \square & \dots & \square & \square & \square \\
 +c & \square & +a & \square & \square & \lambda + \mu & \square & \dots & \square & \square & \square \\
 \square & +c & \square & \square & \square & \square & \lambda + \mu & \dots & \square & \square & \square \\
 \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
 \square & \square & \square & \square & \square & \square & \square & \dots & \mu & \square & \square \\
 \square & \square & \square & \square & \square & \square & \square & \dots & \square & \mu & \square
 \end{bmatrix}$$

Figura 4.9 – Matriz gerada pelos procedimentos de generalização das equações de balanço da fila.

No próximo capítulo encontra-se um estudo de caso no SAMU-RP, utilizando o modelo hipercubo com prioridade na fila. Uma das principais vantagens dessa abordagem é o cálculo das medidas de desempenho para cada classe de usuários, possibilitando avaliar e melhorar o sistema do ponto de vista de alguma classe de usuário.

5 Aplicação do modelo hipercubo no SAMU-RP e análise dos resultados

Neste capítulo, é feita a validação das nove hipóteses descritas na Seção 3.5 do Capítulo 3 e a aplicação e análise do modelo hipercubo estendido com prioridade na fila. As medidas de desempenho são obtidas para o sistema SAMU-RP, investigando-se três períodos do dia de operação do sistema. Além disso, são avaliados cenários alternativos, obtidos com variações das taxas de chegada e nas configurações das localizações das ambulâncias no SAMU-RP.

5.1 Descrição dos atendimentos

O SAMU-RP disponibilizou as fichas de controle das chamadas e de controle de tráfego realizadas no ano de 2005, auxiliando no processo de coleta de dados (Anexo A). A coleta de dados foi feita em duas etapas. Na primeira etapa, o objetivo foi a verificação de possíveis períodos de pico ao longo do ano, meses, semanas e dias. Assim, a coleta de dados foi feita verificando a quantidade de ocorrências de urgência e emergência deste ano ao longo dos meses e também de dez dias do mês de agosto, essa escolha foi feita aleatoriamente. Para estudar a quantidade dos chamados ao longo do tempo, foi feito o levantamento da frequência dos chamados somente por meio das fichas de controle dos chamados.

a) Observações ao longo do ano de 2005.

Foi feito um levantamento do número de atendimentos ao longo do ano, a fim de verificar se há diferenças no número de solicitações, por exemplo, em períodos de férias escolares. Assim, para verificar as possíveis variações da frequência dos chamados, foi construída a distribuição do número de atendimentos realizados em 2005 do SAMU-RP. A Figura 5.1 mostra a distribuição dos chamados de janeiro a dezembro de 2005.

Para verificar se o número médio de atendimentos é igual em todos os

meses de 2005, foi aplicada a Análise de Variância (ANOVA). O resultado mostrou que com $\alpha = 5\%$ de significância, pode-se rejeitar a hipótese de igualdade das médias. A partir do Teste de Tukey [mais detalhes da ANOVA e do teste de Tukey podem ser vistos no Anexo B, em Costa Neto (1977) e em Magalhães e Lima (2002)], foi verificado que apenas o mês de abril tem média diferente dos outros meses. Em consulta ao calendário do ano e ao coordenador do SAMU, esse mês foi considerado atípico, uma vez que houve dois feriados prolongados, o dia 15 (sexta-feira, Paixão) e 21 (quinta-feira, Tiradentes) que ocorreram nesse mês em 2005, que variam de ano para ano e aumentam o número de ocorrências.

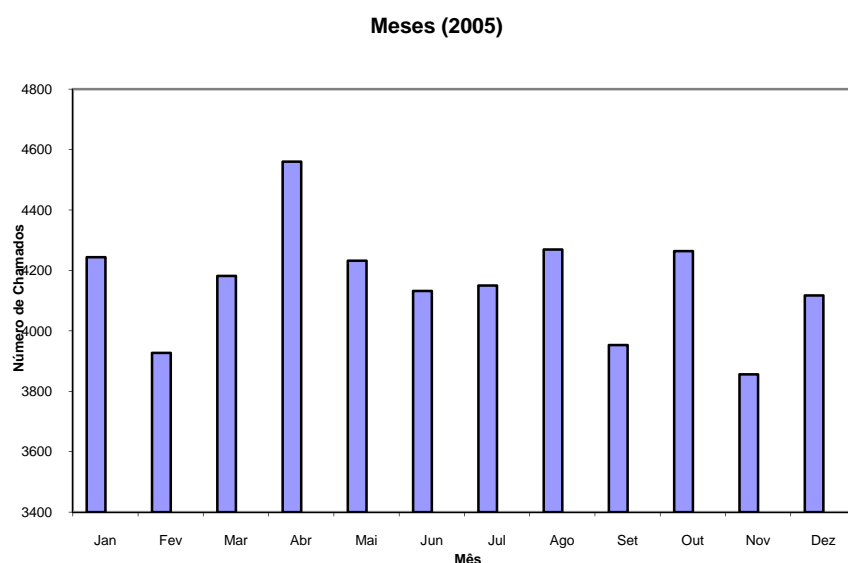


Figura 5.1 – Atendimentos de urgência e emergência realizados pelo SAMU_RP em 2005.

b) Observação ao longo dos meses

Também verificou-se se há diferenças significativas com relação ao número de atendimentos dentro dos meses (por exemplo, se nos finais de semana há mais ou menos atendimentos em relação aos dias úteis). Foi feito um levantamento da frequência dos chamados para cada mês de 2005. A Figura 5.2 mostra a distribuição de frequências do número de atendimentos, por exemplo, do mês de agosto. Observa-se que não há diferenças significativas ao longo desse mês. Este comportamento também foi verificado em todos os meses de 2005.

c) Observações ao longo de dez dias do mês de agosto de 2005.

Dez dias do mês de agosto de 2005 foram escolhidos, aleatoriamente, para

verificar diferenças com relação ao número de chamados ao longo do dia. Com base nessa análise, foram determinados os períodos de pico dos dias analisados neste estudo. Foi feita a análise do número médio de atendimentos dos dez dias de agosto de 2005, por meio da ANOVA. O resultado mostrou que não há diferenças significativas do número médio de chamados, com $\alpha = 5\%$ de significância (ver Anexo B).

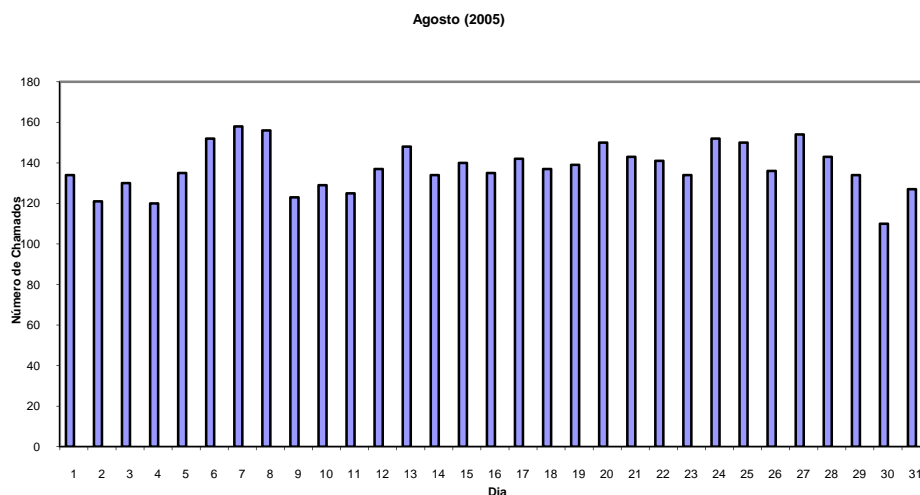


Figura 5.2 – Atendimentos de urgência e emergência realizados pelo SAMU_RP em agosto de 2005.

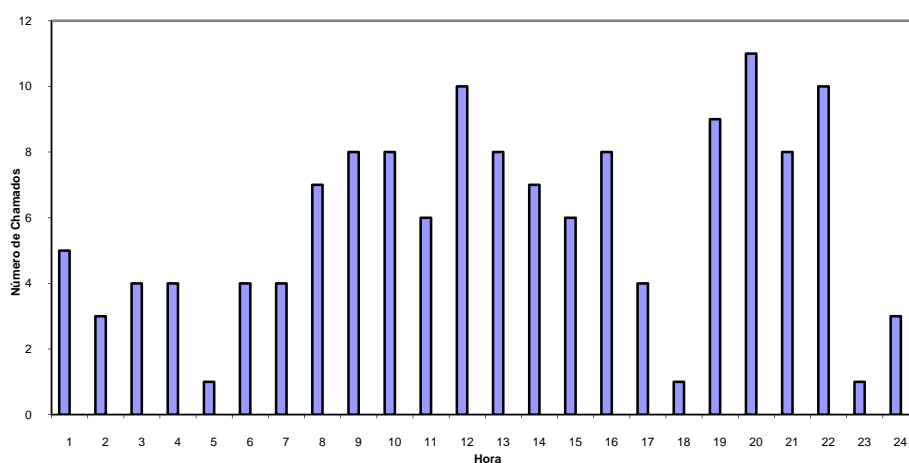


Figura 5.3 – Atendimentos realizados pelo SAMU_RP no 4º dia observado em agosto de 2005.

A Figura 5.3 mostra a distribuição dos chamados no 4º dia observado em agosto de 2005 (por hora). Pode-se verificar que há dois períodos de pico distintos: por volta de 12:00h e 20:00h. Os demais gráficos dos 10 dias observados podem ser vistos no Anexo C.

5.2 Períodos de pico

Um dos objetivos deste trabalho é estudar a configuração do SAMU-RP com múltiplas configurações. Sendo assim, o dia foi dividido em três partes: manhã (08:00h às 16:00h), tarde (16:00h às 24:00h) e noite (00:00h às 08:00h). Para cada parte do dia, foi determinado um período de pico a ser estudado, conforme descrito a seguir.

A análise dos períodos de pico foi feita com os dados obtidos a partir do controle dos chamados do SAMU-RP. A Tabela 5.1 mostra a taxa de chegada (número de chamados dividido pelo período de tempo considerado, em minutos) para vários períodos do dia, durante os 10 dias estudados. Foram analisados vários períodos e o período de pico em cada parte do dia foi identificado pela análise feita a partir da média e do desvio padrão da taxa de chegada. Os períodos escolhidos (linhas destacadas da Tabela 5.1) foram os que apresentaram maior taxa média de chegada e menor desvio-padrão, conforme o estudo em Takeda (2000) para o SAMU-Campinas. Alguns períodos apresentaram a média e o desvio-padrão muito próximos, especialmente no período da noite. Assim, para aqueles períodos com taxa média de chegada pequenos e muito próximos, escolhe-se aquele de maior duração, a fim de obter uma maior quantidade de dados no período. A seguir, são mencionados apenas períodos manhã, tarde e noite, fazendo referência aos três períodos de pico correspondentes.

Dia

Manhã													
Período (horas)		1	2	3	4	5	6	7	8	9	10	Média	D-P
8 - 14		0,1	0,1	0,11	0,1	0,13	0,12	0,1	0,13	0,11	0,11	0,111	0,01156
9 - 14		0,1	0,1	0,11	0,09	0,13	0,11	0,08	0,13	0,12	0,11	0,110	0,01519
10 - 14		0,12	0,12	0,11	0,1	0,13	0,12	0,08	0,12	0,11	0,11	0,113	0,01242
11 - 14		0,11	0,11	0,08	0,11	0,14	0,14	0,09	0,11	0,12	0,11	0,111	0,01842
12 - 14		0,12	0,12	0,08	0,09	0,13	0,12	0,13	0,1	0,12	0,08	0,107	0,01792

Tarde

Período (horas)			1	2	3	4	5	6	7	8	9	10	Média	D-P
20	-	24	0,1	0,13	0,13	0,09	0,1	0,14	0,13	0,1	0,13	0,13	0,118	0,01861
17	-	23	0,1	0,14	0,13	0,11	0,11	0,14	0,13	0,11	0,13	0,11	0,121	0,01550
18	-	23	0,11	0,13	0,15	0,13	0,12	0,16	1,33	0,1	0,14	0,11	0,249	0,38148
19	-	23	0,11	0,12	0,15	0,13	0,1	0,17	0,14	0,11	0,13	0,13	0,127	0,01956
20	-	23	0,11	0,1	0,14	0,11	0,1	0,16	0,14	0,11	0,12	0,13	0,121	0,02063

Noite

Período (horas)			1	2	3	4	5	6	7	8	9	10	Média	D-P
3	-	5	0,03	0,04	0,07	0,04	0,05	0,04	0,06	0,04	0,04	0,04	0,046	0,00982
24	-	6	0,05	0,06	0,05	0,06	0,08	0,05	0,06	0,08	0,04	0,04	0,055	0,01367
24	-	7	0,05	0,07	0,05	0,06	0,08	0,05	0,05	0,08	0,04	0,04	0,056	0,01543
1	-	6	0,04	0,04	0,05	0,05	0,07	0,04	0,05	0,07	0,04	0,04	0,051	0,01235
2	-	6	0,03	0,04	0,05	0,05	0,06	0,03	0,06	0,06	0,03	0,04	0,045	0,01249

Tabela 5.1– Análise do período de pico para os três períodos do dia: manhã, tarde e noite.

Pode-se observar que, caso a análise fosse feita em apenas um período do dia, o escolhido seria à tarde das 18h às 23h. Mais adiante é mostrado que esse não é o período mais desfavorável do ponto de vista das medidas de desempenho, como *workload* e tempos de resposta, comparado com o período da manhã.

5.3 Validação das hipóteses do modelo hipercubo para o SAMU-RP

Faz-se necessário verificar se o sistema atende às nove hipóteses do modelo hipercubo, descritas na Seção 3.2, considerando todas as características do SAMU-RP, descritas no Capítulo 2.

5.3.1 Área dividida em N_A átomos geográficos

Há várias maneiras de se fazer a representação da área estudada em átomos geográficos, como divisão política, bairros, setores policiais, entre outros. Neste

trabalho, pretende-se utilizar a divisão por setores: Norte, Sul, Leste, Oeste e Central, utilizada no SAMU-RP.

Em Takeda *et al.* (2004, 2007), havia dois tipos de chamadas no SAMU-Campinas: básicas e avançadas. Para fins de modelagem, naquele trabalho, os átomos foram biparticionados. Conforme mencionado antes, o SAMU-RP também possui classes diferenciadas de usuários do sistema, chamadas de classificação por risco e descritas na Seção 2.4. Dessa forma, neste trabalho, cada átomo geográfico (Centro – 1; Norte – 2; Sul – 3; Leste – 4; Oeste – 5) foi dividido em três subátomos (*a*, *b*, *c*), totalizando quinze subátomos no sistema: Centro *a* (1*a*), Centro *b* (1*b*), Centro *c* (1*c*), Norte *a* (2*a*), Norte *b* (2*b*), Norte *c* (2*c*), Sul *a* (3*a*), Sul *b* (3*b*), Sul *c* (3*c*), Leste *a* (4*a*), Leste *b* (4*b*), Leste *c* (4*c*), Oeste *a* (5*a*), Oeste *b* (5*b*), Oeste *c* (5*c*). Assim, são devidamente representadas no modelo as três classes de usuários do SAMU-RP (grave (*a*), moderada (*b*) e leve (*c*), Tabela 5.4).

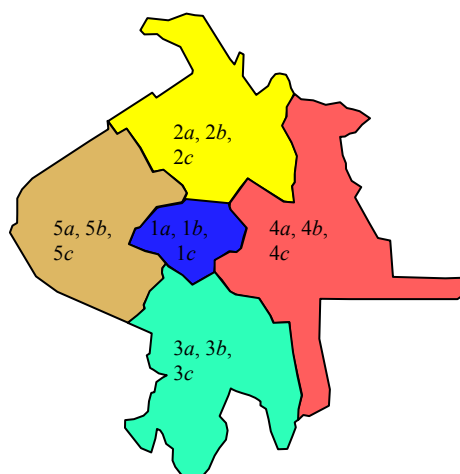


Figura 5.4 – Átomos geográficos do SAMU_RP.

5.3.2 Processo de chegada

Faz-se necessário fazer um teste de aderência nos dados para verificar estatisticamente a hipótese de processo de chegada Poisson. Os métodos utilizados foram *Kolmogorov-Smirnov*, *Anderson Darling* e *Qui-quadrado*; ver Johnson *et. al.* (1994, 1995). Para fazer a análise do processo de chegada nos períodos de pico, foram considerados os chamados nestes períodos, analisando duas fichas: controle dos chamados e controle de tráfego. Os chamados que existiam na ficha de controle dos chamados, mas não puderam ser identificados no controle de tráfego, assim como os

atendimentos de remoção (discutidos na Seção 2.4), foram desconsiderados da análise. Inicialmente, foram considerados dez dias do mês de agosto. Apenas os períodos em que estavam dez ambulâncias trabalhando nos períodos da manhã e da tarde e nove ambulâncias no período da noite foram considerados. Assim, participaram da análise 9, 10 e 6 dias nos períodos da manhã, tarde e noite, respectivamente. A Tabela 5.2 mostra os resultados obtidos do número de chamadas em cada átomo e a proporção (p_j) com relação ao total observado no sistema, no período da manhã.

Subátomos		No de Chamados			Proporção - p_j		
		Manhã	Tarde	Noite	Manhã	Tarde	Noite
1	1a	5	6	1	0,0244	0,0233	0,0088
2	1b	29	30	10	0,1415	0,1167	0,0877
3	1c	15	14	6	0,0732	0,0545	0,0526
4	2a	5	1	1	0,0244	0,0039	0,0088
5	2b	24	26	19	0,1171	0,1012	0,1667
6	2c	13	15	4	0,0634	0,0584	0,0351
7	3a	1	2	0	0,0049	0,0078	0,0000
8	3b	11	9	6	0,0537	0,035	0,0526
9	3c	6	10	8	0,0293	0,0389	0,0702
10	4a	2	1	0	0,0098	0,0039	0,0000
11	4b	25	55	16	0,1220	0,2140	0,1404
12	4c	12	24	10	0,0585	0,0934	0,0877
13	5a	5	4	1	0,0244	0,0156	0,0088
14	5b	24	31	22	0,1171	0,1206	0,1930
15	5c	28	29	10	0,1366	0,1128	0,0877
Total		205	257	114	1,0000	1,0000	1,0000

Tabela 5.2 – Proporção de chamados em cada subátomo do sistema.

Foi feita a análise do processo de chegada para todos os dias de observação, divididos em manhã, tarde e noite (períodos de pico), a fim de verificar se o número de chamadas segue o padrão Poissoniano, uma vez que as chegadas das chamadas em cada átomo constituem processos de contagem com incrementos independentes. Os testes de aderência foram realizados utilizando-se o *software* BestFit. Feitos para os chamados agregados (em todos os átomos), esses testes mostraram que, a um nível de significância de 5%, não se pode rejeitar a hipótese de que os intervalos entre chegadas sucessivas tem distribuição exponencial, nos três períodos do dia. As Figuras 5.5, 5.6 e 5.7 mostram os gráficos obtidos no *software* BestFit do teste de aderência feito com a distribuição Exponencial, para os períodos da manhã, tarde e noite, respectivamente.

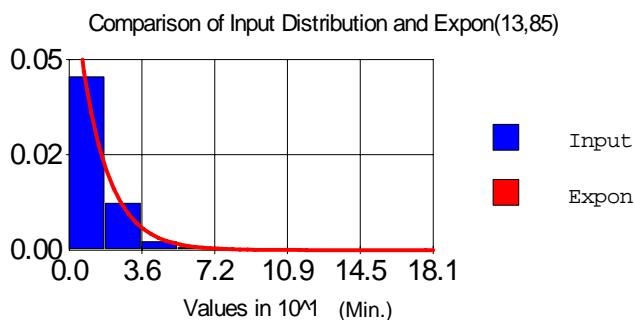


Figura 5.5 - Gráfico obtido no BestFit do teste de aderência do processo de chegada, para o período da manhã.

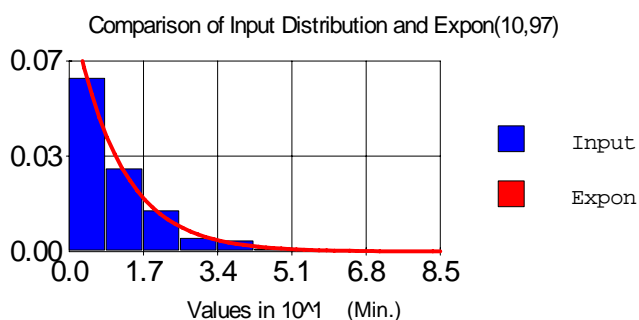


Figura 5.6 - Gráfico obtido no BestFit do teste de aderência do processo de chegada, para o período da tarde.

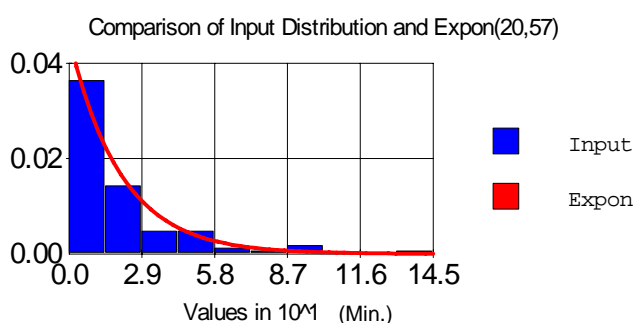


Figura 5.7 - Gráfico obtido no BestFit do teste de aderência do processo de chegada, para o período da noite.

O gráfico ilustrado na Figura 5.8 mostra a porcentagem do número de chamados em cada subátomo nos três períodos: manhã, tarde e noite. Pode-se observar que há diferenças significativas na porcentagem do número de chamados durante os períodos manhã, tarde e noite, principalmente nos subátomos Nb, Sb, Cb e Ob.

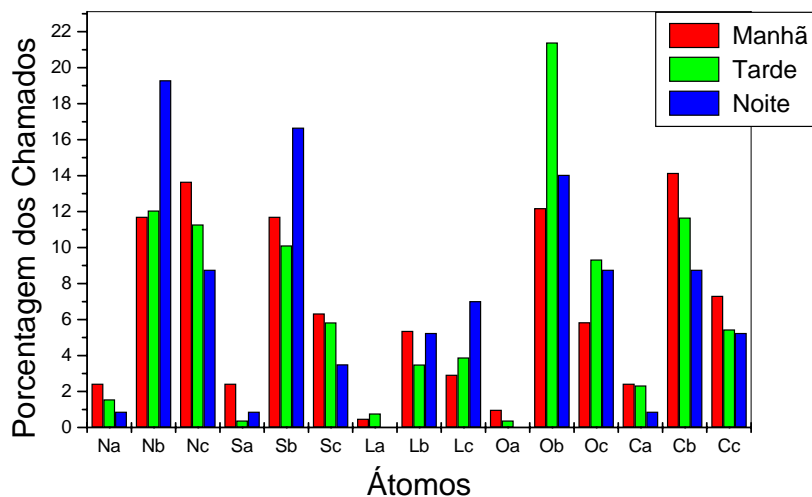


Figura 5.8 - Porcentagem do número de chamados em cada subátomo nos períodos manhã, tarde e noite.

5.3.2.1 Cálculo dos λ_{jk}

A fim de determinar a taxa média de chegada dos chamados no sistema, foram considerados os intervalos médios de chegadas sucessivas para todos os dias de observação nos períodos de pico (manhã, tarde e noite). A Tabela 5.3 apresenta o tempo médio, desvio-padrão e coeficiente de variação dos intervalos entre chamadas para cada dia, nos períodos da manhã, tarde e noite.

Dia	Intervalo médio (min)			Desvio-padrão (min)			Coeficiente de variação		
	Manhã	Tarde	Noite	Manhã	Tarde	Noite	Manhã	Tarde	Noite
1	10,4	11,6	18,0	10,8	17,3	18,3	1,0	1,5	1,0
2	11,7	9,2	22,4	7,6	6,6	25,3	0,7	0,7	1,1
3	10,9	8,4	16,9	12,1	7,1	30,0	1,1	0,8	1,8
4	16,8	8,5	18,1	11,6	9,9	18,8	0,7	1,2	1,0
5	9,3	12,0	33,7	8,5	8,8	28,1	0,9	0,7	0,8
6	30,3	14,0	23,6	33,9	11,3	24,5	1,1	0,8	1,0
7	13,3	13,0	-----	12,8	11,1	-----	1,0	0,9	-----
8	12,9	10,8	-----	14,4	13,0	-----	1,1	1,2	-----
9	13,3	16,8	-----	10,1	13,0	-----	0,8	0,8	-----
10	-----	9,5	-----	-----	12,9	-----	-----	1,4	-----

Tabela 5.3 – Intervalos médios entre chegadas sucessivas no sistema, no período da noite.

Pode-se verificar que em todos os casos, nos três períodos, os desvios-padrão são, em geral, da ordem de grandeza das médias, ou seja, os coeficientes de variação são relativamente próximos de 1. Apenas no período da noite e no 3º dia de observação o intervalo médio entre chegadas é 1,8 (significativamente maior que 1). Isso é mais um indicativo que o intervalo de tempo entre chegadas sucessivas dos chamados deve ser de fato exponencialmente distribuído.

Como visto na seção anterior, é razoável admitir que os chamados agregados chegam segundo Processos de Poisson. Dessa forma, considerando as devidas proporções que representam as chegadas em cada subátomo do sistema, admitindo que os chamados chegam independentemente e de acordo com o Processo de Poisson, encontraram-se as taxas médias $\lambda_{jk} = \lambda \cdot p_{jk}$ ($j = 1, 2, \dots, 5, k \in C$). Pode-se ver na Tabela 5.4 as taxas de chegadas dos chamados considerando as devidas proporções que representam as chegadas dos chamados em cada subátomo do sistema, nos três períodos analisados: manhã, tarde e noite. Em cada período, temos um comportamento diferente da distribuição dos chamados em cada subátomo do sistema.

	Subátomo	Nº de Chamados			Proporção - p_{jk}			λ_{jk}		
		manhã	tarde	noite	manhã	tarde	noite	manhã	tarde	noite
C	1a	5	4	1	0,0244	0,0156	0,0088	0,1164	0,0851	0,0256
	1b	24	31	22	0,1171	0,1206	0,193	0,5586	0,6596	0,5628
	1c	28	29	10	0,1366	0,1128	0,0877	0,6517	0,6171	0,2558
N	2a	5	6	1	0,0244	0,0233	0,0088	0,1164	0,1277	0,0256
	2b	29	30	10	0,1415	0,1167	0,0877	0,675	0,6384	0,2558
	2c	15	14	6	0,0732	0,0545	0,0526	0,3491	0,2979	0,1535
S	3a	5	1	1	0,0244	0,0039	0,0088	0,1164	0,0213	0,0256
	3b	24	26	19	0,1171	0,1012	0,1667	0,5586	0,5532	0,486
	3c	13	15	4	0,0634	0,0584	0,0351	0,3026	0,3192	0,1023
L	4a	1	2	0	0,0049	0,0078	0,0000	0,0233	0,0426	0,0000
	4b	11	9	6	0,0537	0,035	0,0526	0,256	0,1915	0,1535
	4c	6	10	8	0,0293	0,0389	0,0702	0,1397	0,2128	0,2047
O	5a	2	1	0	0,0098	0,0039	0,0000	0,0466	0,0213	0,0000
	5b	25	55	16	0,122	0,214	0,1404	0,5819	1,1703	0,4093
	5c	12	24	10	0,0585	0,0934	0,0877	0,2793	0,5107	0,2558
Total		205	257	114	1	1	1	4,7716	5,4686	2,9163

Tabela 5.4 – Taxas médias de chegada dos chamados (por hora) para cada subátomo.

5.3.3 Tempos de viagem

Faz-se necessário calcular os tempos médios de viagem de cada ambulância para cada átomo, podendo ser obtidos no próprio sistema. Caso não haja dados suficientes, os tempos de viagem entre os átomos podem ser calculados a partir de conceitos de probabilidade geométrica (LARSON e ODONI, 2007). Os tempos de viagem foram obtidos a partir dos dados do próprio SAMU-RP. Apenas em um caso não foi encontrado observações do tempo de viagem entre dois átomos. Nesse caso, foi calculada a distância entre os centróides dos átomos (a partir do *software* Google Earth) e, utilizando a velocidade média de 60 km/h, foi possível obter uma estimativa do tempo médio de viagem entre os átomos (indicados na tabela a seguir com asterisco “*”). A matriz dos tempos de viagem entre todos os subátomos, para os três períodos, pode ser vista na Tabela 5.5.

	1a	1b	1c	2a	2b	2c	3a	3b	3c	4a	4b	4c	5a	5b	5c
1a	8,7	8,7	8,7	9,5	9,5	9,5	10,9	10,9	10,9	11,4	11,4	11,4	9,9	9,9	9,9
1b	8,7	8,7	8,7	9,5	9,5	9,5	10,9	10,9	10,9	11,4	11,4	11,4	9,9	9,9	9,9
1c	8,7	8,7	8,7	9,5	9,5	9,5	10,9	10,9	10,9	11,4	11,4	11,4	9,9	9,9	9,9
2a	9,5	9,5	9,5	11,6	11,6	11,6	9,7	9,7	9,7	7,3	7,3	7,3	7,5	7,5	7,5
2b	9,5	9,5	9,5	11,6	11,6	11,6	9,7	9,7	9,7	7,3	7,3	7,3	7,5	7,5	7,5
2c	9,5	9,5	9,5	11,6	11,6	11,6	9,7	9,7	9,7	7,3	7,3	7,3	7,5	7,5	7,5
3a	10,9	10,9	10,9	9,7	9,7	9,7	8,6	8,6	8,6	13,5	13,5	13,5	10,7	10,7	10,7
3b	10,9	10,9	10,9	9,7	9,7	9,7	8,6	8,6	8,6	13,5	13,5	13,5	10,7	10,7	10,7
3c	10,9	10,9	10,9	9,7	9,7	9,7	8,6	8,6	8,6	13,5	13,5	13,5	10,7	10,7	10,7
4a	11,4	11,4	11,4	7,3	7,3	7,3	13,5	13,5	13,5	10,1	10,1	10,1	15,1*	15,1*	15,1*
4b	11,4	11,4	11,4	7,3	7,3	7,3	13,5	13,5	13,5	10,1	10,1	10,1	15,1*	15,1*	15,1*
4c	11,4	11,4	11,4	7,3	7,3	7,3	13,5	13,5	13,5	10,1	10,1	10,1	15,1*	15,1*	15,1*
5a	9,9	9,9	9,9	7,5	7,5	7,5	10,7	10,7	10,7	15,1*	15,1*	15,1*	9,2	9,2	9,2
5b	9,9	9,9	9,9	7,5	7,5	7,5	10,7	10,7	10,7	15,1*	15,1*	15,1*	9,2	9,2	9,2
5c	9,9	9,9	9,9	7,5	7,5	7,5	10,7	10,7	10,7	15,1*	15,1*	15,1*	9,2	9,2	9,2

Tabela 5.5 – Tempo médio de viagem entre subátomos obtidos a partir de dados do SAMU-RP.

5.3.4 Servidores

O sistema do SAMU-RP é composto por uma frota de dez ambulâncias distintas: nove VSB's e um VSA. As ambulâncias estão descentralizadas, localizadas em cinco postos de saúde distribuídos um em cada setor da cidade, podendo deslocar-se

para qualquer átomo para realizar um atendimento. As ambulâncias estão distribuídas da seguinte forma, dois VSB's e um VSA na região central, a Região Norte, as Regiões Sul e Oeste possuem dois VSB's cada e a Região Leste possui uma VSB, nos períodos da manhã e tarde. À noite, a região Oeste trabalha com uma VSB, totalizando 9 ambulâncias, neste período.

As Ambulâncias VSB's possuem um motorista, que pode participar do atendimento, e um auxiliar de enfermagem, enquanto a ambulância VSA, também conhecida como UTI-Móvel, possui equipamentos de primeiros socorros, como medicamentos, oxigênio, ressuscitadores, macas e outros. A VSA também conta com uma equipe formada por médico, enfermeira e motorista (podendo auxiliar no atendimento, caso necessário).

5.3.5 Localização dos servidores

A matriz de localização (L) é obtida a partir da configuração original do sistema, de acordo com os critérios utilizados no SAMU-RP. Para generalizar a nomenclatura, o VSA passa a ser chamado de veículo 1, enquanto os VSB's passam a ser chamados de 2, 3, 4, 5, 6, 7, 8, 9 e 10, para os períodos da manhã e tarde. A matriz de localização para os períodos da manhã e tarde pode ser vista na Tabela 5.6. No período noturno há nove ambulâncias operando. A matriz de localização para o período da noite pode ser vista na Tabela 5.7.

	1a	1b	1c	2a	2b	2c	3a	3b	3c	4a	4b	4c	5a	5b	5c	Amb.
$L =$	1/3	1/3	1/3	0	0	0	0	0	0	0	0	0	0	0	0	1
	1/3	1/3	1/3	0	0	0	0	0	0	0	0	0	0	0	0	2
	1/3	1/3	1/3	0	0	0	0	0	0	0	0	0	0	0	0	3
	0	0	0	1/3	1/3	1/3	0	0	0	0	0	0	0	0	0	4
	0	0	0	1/3	1/3	1/3	0	0	0	0	0	0	0	0	0	5
	0	0	0	0	0	0	1/3	1/3	1/3	0	0	0	0	0	0	6
	0	0	0	0	0	0	1/3	1/3	1/3	0	0	0	0	0	0	7
	0	0	0	0	0	0	0	0	0	1/3	1/3	1/3	0	0	0	8
	0	0	0	0	0	0	0	0	0	0	0	0	1/3	1/3	1/3	9
	0	0	0	0	0	0	0	0	0	0	0	0	1/3	1/3	1/3	10

Tabela 5.6 – Matriz de localização dos servidores nos subátomos para os períodos da manhã e tarde.

	1a	1b	1c	2a	2b	2c	3a	3b	3c	4a	4b	4c	5a	5b	5c	Amb.
$L =$	1/3	1/3	1/3	0	0	0	0	0	0	0	0	0	0	0	0	1
	1/3	1/3	1/3	0	0	0	0	0	0	0	0	0	0	0	0	2
	1/3	1/3	1/3	0	0	0	0	0	0	0	0	0	0	0	0	3
	0	0	0	1/3	1/3	1/3	0	0	0	0	0	0	0	0	0	4
	0	0	0	1/3	1/3	1/3	0	0	0	0	0	0	0	0	0	5
	0	0	0	0	0	0	1/3	1/3	1/3	0	0	0	0	0	0	6
	0	0	0	0	0	0	1/3	1/3	1/3	0	0	0	0	0	0	7
	0	0	0	0	0	0	0	0	0	1/3	1/3	1/3	0	0	0	8
	0	0	0	0	0	0	0	0	0	0	0	0	1/3	1/3	1/3	9

Tabela 5.7 – Matriz de localização dos servidores nos subátomos para o período da noite.

5.3.6 Despacho dos servidores

O SAMU-RP admite enviar apenas uma equipe para atender um chamado, atendendo um dos requisitos para a aplicação do modelo hipercubo. O VSA atende apenas chamados de emergência, caracterizando *backup* parcial. No modelo proposto, foi simplificada esta característica do sistema, considerando que todos os servidores atendem a todos os tipos de chamados. Essa simplificação é razoável, uma vez que as frequências de despacho do VSA para chamados do tipo *b* e *c* são bem pequenas, i.e: 0,0073, 0,0023 e menores que 10^{-4} , para os períodos da manhã, tarde e noite, respectivamente.

A formação de fila é permitida quando os usuários solicitam atendimento enquanto todas as ambulâncias estão ocupadas. Para fins de modelagem, a fila foi limitada em 5 usuários. Porém, o SAMU-RP opera sem limite para fila (i.e., fila infinita). Essa simplificação é razoável, uma vez que a probabilidade de perda no modelo é menor que 10^{-4} para os três períodos analisados.

A fila de espera é formada e a escolha do próximo usuário a ser atendido é feita a partir da prioridade do chamado, na ordem: grave, moderado e leve. Observando-se ainda, que o VSA atende apenas a chamados graves, enquanto os VSB's atendem a qualquer tipo de chamado.

5.3.7 Política de despacho dos servidores

Ao receber um chamado e identificar a gravidade do caso, urgência (prioridade A) ou emergência (prioridades B ou C), o médico regulador decide se um VSA ou um VSB atenderá o chamado e entra em contato com os hospitais do município para verificar a disponibilidade de vagas para o caso em questão.

A política de despacho dos servidores depende da distribuição espacial e da localização dos servidores. Lembrando que, na cidade de Ribeirão Preto, os servidores estão descentralizados, a escolha do servidor preferencial é feita analisando a origem e a gravidade do chamado, e a preferência é dada ao servidor localizado na mesma área (escolhido aleatoriamente). Se todos os servidores da área do chamado estiverem ocupados, é escolhido o primeiro servidor disponível mais próximo do chamado. Se o chamado for urgente, um VSA é imediatamente enviado ao local da ocorrência. Caso o VSA esteja ocupado, o VSB mais próximo do local do chamado é enviado. Se o chamado for uma emergência, o VSB disponível mais próximo do local é enviado (o VSA não atende emergências).

Caso todos os servidores estiverem ocupados, o chamado entra em uma fila de espera. O chamado com maior prioridade da fila é atendido pelo primeiro VSB disponível mais próximo da ocorrência. Tanto o VSA quanto os VSB's podem se deslocar para qualquer átomo. A ordem de prioridade para os chamados está descrita na Seção 2.4.

A lista de despacho em que um átomo tem mais de um servidor preferencial, ou seja, casos de desempate de prioridade entre ambulâncias de mesmo local, podem ser incorporadas no modelo. Isso pode ser feito pela introdução da distribuição de frequências de despachos de cada servidor para cada átomo nas equações de balanço do sistema, ou considerando um número suficientemente grande de listas de preferências de despacho geradas aleatoriamente, representando, dessa forma, as possíveis chances dos servidores primários (e/ou *backup*) de cada átomo serem enviados para atender um chamado em cada cenário investigado (BURWELL, *et al.* 1993; TAKEDA, 2000).

O método escolhido foi da geração aleatória da política de despacho dos servidores, sugerido por Burwell *et.al.* (1993). Nos átomos Central, Norte, Sul e Oeste há duas ambulâncias VSB's. Elas são escolhidas preferencialmente e, com a mesma

chance, para chamados b ou c , ocupando a 1ª e 2ª preferências. Caso as duas ambulâncias estejam ocupadas, qualquer ambulância VSB pode atender ao chamado. Dessa forma, é feito um sorteio para a escolha das próximas preferências, quando todas as ambulâncias VSB's restantes têm a mesma chance de serem escolhidas em todas as posições. O VSA atende a estes chamados somente se todas as ambulâncias VSB's estiverem ocupadas. Na região Leste, há apenas uma ambulância: dessa forma, ela é sempre a primeira escolha para chamados b ou c e o procedimento para a escolha das próximas posições é semelhante aos átomos com 2 ambulâncias.

Quando surge um chamado com prioridade a em qualquer átomo, o VSA é sempre o primeiro a ser escolhido. Caso o VSA esteja ocupado, a escolha para as próximas preferências dos VSB's é feita considerando as localizações de cada ambulância, similarmente nos átomos com prioridades b ou c . Uma das possíveis matrizes obtidas é mostrada na Tabela 5.8. Para que essa política de despacho tenha efeito, o modelo hipercubo deve ser executado várias vezes (replicações), as probabilidades de estado finais são as médias de todas as probabilidades de estado calculadas em cada replicação executada. No modelo hipercubo adaptado, foram feitos vários testes e verificou-se que com 20 replicações as probabilidades de estado convergiam.

Subátomo	Preferências de despacho									
	1ª	2ª	3ª	4ª	5ª	6ª	7ª	8ª	9ª	10ª
1 - 1a	1	2	3	9	10	5	4	6	8	7
2 - 1b	2	3	4	9	5	10	7	8	6	1
3 - 1c	3	2	10	9	7	4	6	8	5	1
4 - 2a	1	4	5	9	6	8	2	3	7	10
5 - 2b	5	4	9	7	6	3	10	2	8	1
6 - 2c	5	4	2	6	8	9	3	7	10	1
7 - 3a	1	6	7	8	3	9	5	4	2	10
8 - 3b	7	6	8	2	10	3	9	5	4	1
9 - 3c	6	7	3	2	5	4	9	8	10	1
10 - 4a	1	8	7	5	10	4	6	2	9	3
11 - 4b	8	3	9	6	5	2	7	4	10	1
12 - 4c	8	9	5	6	3	7	2	10	4	1
13 - 5a	1	10	9	2	3	5	7	6	8	4
14 - 5b	9	10	6	5	8	3	2	4	7	1
15 - 5c	9	10	7	2	6	3	8	4	5	1

Tabela 5.8 – Exemplo: matriz de preferência de despachos no cenário original.

5.3.8 Tempos de atendimento

Os tempos de atendimento são obtidos a partir do intervalo de tempo entre os instantes de saída e retorno à base. Esses valores estão disponíveis nas fichas de regulação médica. O tempo de atendimento é a soma dos tempos de preparo da equipe, viagem até o local da ocorrência (ida), em cena e viagem de retorno à base (volta). Incluem, portanto: o tempo de viagem, definido como a saída da base até a chegada no local; o tempo em cena, desde o momento em que a ambulância chega no local onde a equipe realiza o atendimento às vítimas, quando for o caso, até o momento que ela sai do local; o tempo de viagem de volta, que é o tempo desde o instante de saída do local até o momento em que a ambulância retorna à base.

Para verificar estatisticamente a hipótese de que os tempos de atendimento são exponencialmente distribuídos, foi usado o teste de Kolmogorov-Smirnov no *software* BestFit. Em todas as ambulâncias, nos três períodos do dia, foi rejeitada a hipótese de que os tempos de serviço são exponencialmente distribuídos, com $\alpha = 5\%$ de significância.

A Tabela 5.9 mostra os tempos médios de atendimento (em minutos), desvios-padrão, coeficiente de variação e as taxas médias para cada ambulância, em cada período do dia: manhã, tarde e noite, respectivamente. Para verificar a hipótese de diferenças nos tempos de atendimento entre os servidores, foi realizada a análise de variância ANOVA (COSTA NETO, 1977; MAGALHÃES e LIMA, 2002) com nível de significância $\alpha = 5\%$. Os resultados mostraram que as diferenças entre as médias dos tempos de atendimento entre os servidores são significativas nos três períodos. Dessa forma, a aplicação do modelo hipercubo deve considerar que os servidores não são homogêneos para os três períodos considerados. De acordo com Larson (1974, 2007) e Jarvis (1985), este tipo de sistema pode ser analisado aproximadamente pelo modelo hipercubo sem que a análise seja muito comprometida. Esta aproximação também foi feita em Takeda (2007), Iannoni (2005) e Iannoni *et. al.* (2006, 2008a, 2008b) sem que a análise dos sistemas fosse comprometida.

Nos três períodos, o tempo médio de atendimento é diferente do desvio-padrão, ou seja, o coeficiente de variação é bem menor que 1, indicando que os tempos de atendimento não são exponencialmente distribuídos. Um experimento com simulação também foi realizado para verificar a validade de se admitir distribuição exponencial nos tempos de serviço dos servidores. Os resultados desse experimento são discutidos adiante na Seção 5.4.1.

Ambulância	Tempo Médio de Serviço (min.)			Desvio Padrão			Coeficiente de Variação			μ (chamados/hora)			Tempo Médio de Viagem (min.)			Desvio Padrão		
	manhã	tarde	noite	manhã	tarde	noite	manhã	tarde	noite	manhã	tarde	noite	manhã	tarde	noite	manhã	tarde	noite
1	41,09	47,17	34,25	12,43	14,12	12,34	0,30	0,30	0,36	1,4602	1,2721	1,7518	10,13	9,83	10,00	3,91	3,37	5,66
2	44,05	30,40	20,95	26,18	18,75	7,82	0,59	0,62	0,37	1,3622	1,9740	2,8643	9,54	8,77	8,36	3,51	4,63	3,34
3	44,05	30,40	20,95	26,18	18,75	7,82	0,59	0,62	0,37	1,3622	1,9740	2,8643	9,54	8,77	8,36	3,51	4,63	3,34
4	60,79	45,58	45,03	36,17	28,70	33,14	0,60	0,63	0,74	0,9871	1,3165	1,3323	12,31	11,73	10,92	8,03	5,72	4,92
5	60,79	45,58	45,03	36,17	28,70	33,14	0,60	0,63	0,74	0,9871	1,3165	1,3323	12,31	11,73	10,92	8,03	5,72	4,92
6	51,39	33,09	26,53	40,30	13,58	9,78	0,78	0,41	0,37	1,1674	1,8130	2,2619	9,32	9,74	9,13	6,09	5,11	2,92
7	51,39	33,09	26,53	40,30	13,58	9,78	0,78	0,41	0,37	1,1674	1,8130	2,2619	9,32	9,74	9,13	6,09	5,11	2,92
8	79,64	38,25	35,62	58,39	18,88	29,08	0,73	0,49	0,82	0,7534	1,5686	1,6847	12,09	10,19	10,00	3,81	5,61	5,12
9	46,53	39,63	25,59	31,16	30,03	10,95	0,67	0,76	0,43	1,2895	1,5141	2,3444	8,97	8,96	8,33	6,41	5,31	3,41
10	46,53	39,63	...	31,16	30,03	...	0,67	0,76	...	1,2895	1,5141	...	8,97	8,96	...	6,41	5,31	...
VSA	41,09	47,17	34,25	12,43	14,12	12,34	0,30	0,30	0,36	1,4602	1,2721	1,7518	10,13	9,83	10,00	3,91	3,37	5,66
VSB	53,91	37,29	30,78	36,22	22,33	17,69	0,67	0,59	0,53	1,1517	1,6449	2,1183	10,26	9,84	9,39	5,76	5,24	3,86
Total	52,62	38,28	31,16	33,84	21,51	17,09	0,63	0,56	0,51	1,1826	1,6076	2,0776	10,25	9,84	9,46	5,58	5,05	4,06

Tabela 5.9 – Tempos e taxas médias de atendimento para cada ambulância.

5.3.9 Relação entre o tempo de atendimento e o tempo de viagem

De acordo com a hipótese 9 do modelo hipercubo, é necessário verificar se os tempos médios de viagem são pequenos em relação aos tempos médios de atendimento para cada ambulância. Em alguns sistemas, pode ocorrer que os tempos de viagem representem uma importante parcela no cálculo dos tempos médios de atendimento. Por exemplo, em sistemas em que a ambulância atende uma chamada em área rural, os tempos de viagem da base ao local da chamada são parcelas relevantes a serem consideradas. Se isso ocorrer, é necessário fazer um processo de calibração dos tempos médios de atendimento, ajustando os tempos médios de viagem separadamente de cada servidor, de forma a considerar os fatores geográficos que influenciam a viagem de cada servidor. Em Larson e Odoni (2007) há uma descrição de um processo iterativo para calibrar μ^{-1} , a partir dos tempos médios de viagem (\overline{TU}_j). O procedimento consiste em verificar a diferença entre os tempos que compõem μ^{-1} , utilizados como entrada no modelo. Caso a diferença seja significativa, o modelo deve ser rodado novamente com os μ^{-1} calculados pelo modelo. Em Chiyoshi *et al.* (2000) há diversos exemplos em que a convergência foi obtida na segunda ou terceira iterações, muito embora não haja prova dessa convergência.

Ambulância	Tempo médio de serviço(min)			Tempo médio de viagem(min)			Relação: tempo viagem/ tempo médio de serviço		
	manhã	tarde	noite	manhã	tarde	noite	manhã	tarde	noite
1	41	47	34	10	10	10	0,25	0,21	0,29
2	44	30	21	10	9	8	0,22	0,29	0,40
3	44	30	21	10	9	8	0,22	0,29	0,40
4	61	46	45	12	12	11	0,20	0,26	0,24
5	61	46	45	12	12	11	0,20	0,26	0,24
6	51	33	27	9	10	9	0,18	0,29	0,34
7	51	33	27	9	10	9	0,18	0,29	0,34
8	80	38	36	12	10	10	0,15	0,27	0,28
9	47	40	26	9	9	8	0,19	0,23	0,33
10	47	40	-----	9	9	-----	0,19	0,23	-----
VSA	41	47	34	10	10	10	0,25	0,21	0,29
VSB	54	37	31	10	10	9	0,19	0,27	0,32
Total	53	38	31	10	10	9	0,20	0,26	0,32

Tabela 5.10 - Relação entre o tempo de atendimento e o tempo de viagem para as ambulâncias.

A Tabela 5.10 mostra o tempo médio de serviço, o tempo médio de viagem e a relação entre o tempo médio de atendimento e o tempo médio de viagem para cada servidor. Pode-se notar que os tempos médios de viagem são relativamente pequenos com relação aos tempos médios de atendimento. Os tempos médios de viagem representam, no máximo, 40% do tempo total de atendimento das ambulâncias 2 e 3 (região central) no período da noite, de forma que a hipótese 9 do modelo hipercubo está validada.

5.4 Resultados da aplicação do modelo hipercubo

O modelo hipercubo estendido para prioridade de fila (conforme descrito no Capítulo 4) foi implementado computacionalmente em linguagem Pascal e executado em um computador com processador Intel® Core™ 2 Due, tecnologia Centrino, 250 GB de HD e 3Gb de RAM em um sistema operacional Windows XP.

O modelo foi aplicado ao SAMU-RP a fim de fazer a análise descritiva do sistema, do ponto de vista das suas medidas de desempenho. Como descrito na Seção 5.2, o dia (24 horas) foi dividido em três períodos: manhã, tarde e noite. Assim, o sistema foi analisado nos três períodos, independentemente, com suas taxas de chegada de cada átomo e término de serviço de cada ambulância. Partiu-se da hipótese de que a distribuição dos chamados e dos serviços pode variar nos diferentes períodos do dia, podendo resultar, para cada período, uma configuração diferente para o sistema.

As equações de equilíbrio são determinadas de forma similar ao exemplo ilustrativo, descrito no Capítulo 4, que representa um sistema de atendimento emergencial utilizando disciplina de prioridade na fila. Dado que o SAMU-RP tem 10 servidores, há $2^{10} = 1024$ estados possíveis no sistema. A matriz de preferência de despacho gerada aleatoriamente (um exemplo é mostrado na Seção 5.3.7) é considerada na geração das equações de equilíbrio do sistema. O sistema de equações da fila é similar no sistema (4.32); no SAMU-RP não há limitações para o tamanho da fila. Porém, na construção do modelo foi considerado que o sistema atende até 5 usuários na fila (gerando um total de 55 estados). Conforme mencionado, essa é uma simplificação razoável, uma vez que a probabilidade de haver mais que 5 usuários em fila é muito pequena, da ordem de 10^{-4} .

O SAMU-RP realiza atendimentos de remoção, que se caracterizam pelo transporte de pacientes entre hospitais, de casa para hospital ou vice-versa, desde que o paciente esteja impossibilitado de locomoção. Esses atendimentos podem ser agendados ou não, e são feitos desde que haja mais que uma ambulância disponível e não ocorram chamados de urgência e/ou emergência na fila. Apesar da baixa prioridade no atendimento, há um grande volume de atendimentos de remoção por mês. São atendidos cerca de 6.000 remoções por mês. Esses chamados geram uma sobrecarga de trabalho nos atendimentos das ambulâncias, pois um atendimento de remoção raramente é interrompido para atender a um chamado com prioridade maior.

A coleta de dados foi feita apenas com chamados de urgência (prioridade *b* ou *c*) e emergência (prioridade *a*). Assim, a primeira análise do SAMU-RP é feita sem considerar atendimentos de remoção e os resultados são mostrados e discutidos a seguir.

5.4.1 Resultados para o SAMU-RP (cenário original)

Os resultados da análise do SAMU-RP foram obtidos por meio da solução, pelo método exato (LARSON, 1974), do sistema de equações de equilíbrio, para cada um dos períodos: manhã, tarde e noite. A Figura 5.9 mostra a localização das ambulâncias nas cinco bases do SAMU-RP, nos períodos da manhã e tarde.

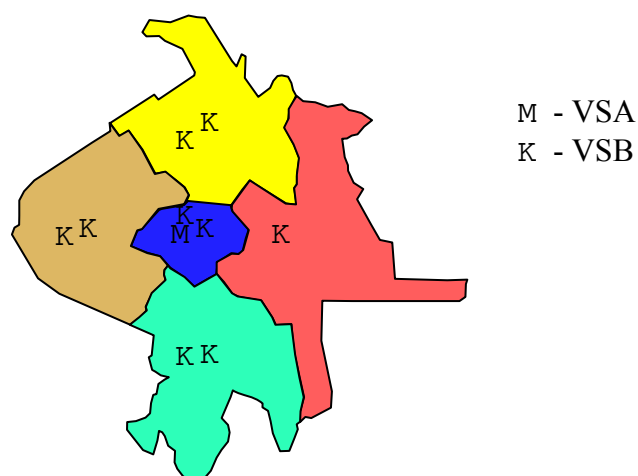


Figura 5.9 – Localização das ambulâncias no SAMU-RP.

A Figura 5.10 mostra o comportamento das *workloads* de todas as ambulâncias, nos três períodos estudados. Nos períodos da manhã e noite, as ambulâncias localizadas na região norte (VSB's 4 e 5) têm *workload* maior que as demais. Esse comportamento se deve ao fato de que as taxas de chegada são maiores na região norte (Tabela 5.4). No período da tarde, são as ambulâncias das regiões norte (VSB's 4 e 5) e oeste (VSB's 9 e 10) que têm *workload* maiores que as demais. Nesse período, as taxas de chegada nessa região são maiores que nas demais regiões. Vale salientar que dados incompletos ou inconsistentes (como por exemplo, tempos de viagem de 1 ou 40 minutos) do controle dos chamados ou do controle de tráfego foram eliminados da amostra, de forma que a comparação dos *workloads* do modelo e da amostra não pôde ser realizada sem prejuízo da análise, uma vez que o *workload* da amostra seria subestimado porque esses períodos em que as ambulâncias estavam ocupadas na amostra não seriam considerados.

A hipótese do modelo hipercubo relativo aos tempos de serviço serem exponencialmente distribuídos foi rejeitada, como mostrado na Seção 5.3.8. A fim de verificar se os resultados do modelo hipercubo são sensíveis a essa hipótese, um modelo de simulação foi desenvolvido e implementado em ARENA utilizando, nos tempos de serviço, as distribuições empíricas obtidas a partir do teste de aderência feito pelo *software* BestFit® (boa parte destas distribuições são lognormais). Os detalhes dessa simulação são semelhantes ao das simulações anteriores, descritos no Anexo D.

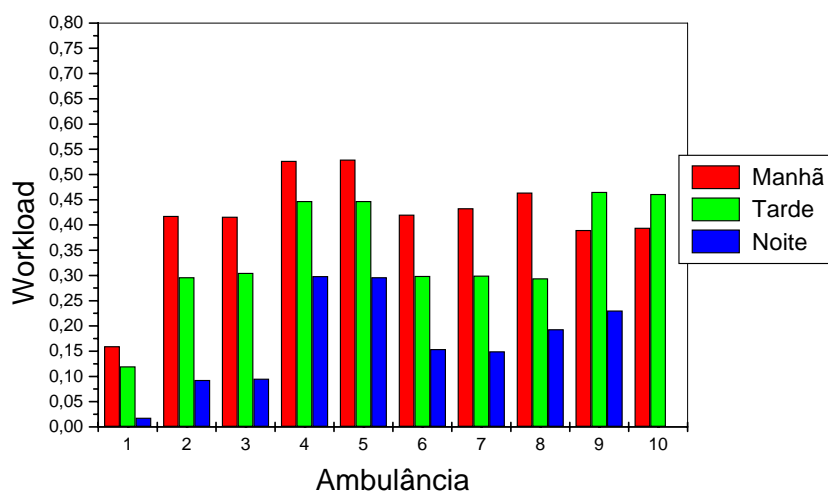


Figura 5.10 – *Workloads* das ambulâncias.

Resultados das medidas de desempenho para o período da tarde.

A probabilidade de encontrar o sistema vazio ($p[0000000000]$) e a probabilidade de todos os servidores estarem ocupados ($P_{1111111111}$) é de 0,0002 e 0,0020, respectivamente. A probabilidade de fila no sistema (P_Q) e o tempo de espera na fila (W_q) é de 0,0009 e 0,0205 minutos, respectivamente.

A Tabela 5.11 mostra os tempos médios de espera em fila do modelo e da simulação. Pode-se notar que os desvios absolutos do modelo em relação à simulação para o tempo médio de espera na fila não são significativos (todos próximos de zero).

Tempo médio de espera em fila (min.)	Sistema	<i>a</i>	<i>b</i>	<i>c</i>
Modelo	0,0100	0,0100	0,0100	0,0100
Simulação	0,0141	0,0105	0,0128	0,0168
Desvio (min.)	-0,0041	-0,0005	-0,0028	-0,0068

Tabela 5.11– Tempo médio de espera na fila.

A Tabela 5.12 mostra os *workloads*, os tempos médios de viagem e de resposta do modelo e da simulação. Pode-se verificar que o desvio da ambulância 1 (VSA) é alto, 35,38%. Isso pode ser devido à pequena quantidade de dados obtidos para essa ambulância (apenas 5 observações), de forma que nenhuma distribuição pode ser bem aproximada pela amostra. A Tabela 8, do Anexo E, mostra os *workloads* correspondentes do período da manhã. Pode-se notar que eles são mais altos do que no período da tarde.

Ambulância	Workload		Desvio	
	Modelo	Simulação	Minutos	%
1	0,1200	0,1942	-0,0742	-38,21
2	0,3000	0,2895	0,0105	3,63
3	0,3100	0,2888	0,0212	7,34
4	0,4500	0,4430	0,0070	1,58
5	0,4500	0,4400	0,0100	2,27
6	0,3000	0,2958	0,0042	1,42
7	0,3000	0,2938	0,0062	2,11
8	0,2900	0,2838	0,0062	2,18
9	0,4700	0,4619	0,0081	1,75
10	0,4600	0,4609	-0,0009	-0,20
VSA	0,1200	0,1942	-0,0742	-38,21
VSB	0,3700	0,3619	0,0081	2,45
Total	0,3450	0,3452	-0,0002	-1,61

Tabela 5.12 – Workload das ambulâncias.

(i) Tempos de Viagem

A Tabela 5.13 mostra o tempo médio de viagem e de resposta no sistema para o modelo e a simulação. Pode-se verificar que os desvios do modelo em relação à simulação também são relativamente pequenos, iguais a 1,5%.

		Tempo médio de viagem no sistema	Tempo médio de resposta no sistema
Modelo		9,7000	9,7100
Simulação		9,7218	9,7359
Desvio	Minutos	-0,0218	-0,0259
	%	-0,22	-0,27

Tabela 5.13 – Tempo médio de viagem e de resposta no sistema.

A Tabela 5.14 mostra os resultados dos tempos médios de viagem para cada ambulância, obtidos pelo modelo, pela simulação e pela amostra. Os desvios do modelo em relação a simulação são relativamente baixos, todos menores que 10%. Os desvios do modelo em relação à amostra também são baixos, o maior deles é 10,8%.

Ambulância	Tempo médio de viagem (minutos)			Desvio em relação a simulação		Desvio em relação a amostra	
	Modelo	Simulação	Amostra	Minutos	%	Minutos	%
1	9,7	9,6	9,8	0,1	1,6	-0,1	-1,0
2	8,8	9,4	8,8	-0,6	-6,3	0,0	0,0
3	8,9	9,4	8,8	-0,5	-5,2	0,1	1,1
4	11,2	10,5	11,7	0,7	6,3	-0,5	-4,3
5	11,2	10,6	11,7	0,6	5,9	-0,5	-4,3
6	9,1	9,6	9,7	-0,5	-5,6	-0,6	-6,2
7	9,5	9,6	9,7	-0,1	-1,4	-0,2	-2,1
8	11,3	11,0	10,2	0,3	3,0	1,1	10,8
9	9,3	9,3	9,0	0,0	-0,5	0,3	3,3
10	9,1	9,4	9,0	-0,3	-2,7	0,1	1,1
VSA	9,7	9,6	9,8	0,1	1,6	-0,1	-1,0
VSB	9,8	9,9	9,8	0,0	-0,7	0,0	0,0
Total	9,8	9,8	9,8	0,0	-0,5	0,0	-0,1

Tabela 5.14 – Tempos médios de viagem das ambulâncias.

A Tabela 5.15 mostra os resultados dos tempos médios de viagem para cada subátomo, obtidos pelo modelo, pela simulação e pela amostra. Apenas no subátomo 3a o desvio foi alto, 16,38%, em geral, os desvios do modelo em relação a simulação são

baixos, todos menores que 10%. Os desvios do modelo em relação à amostra também são baixos, o maior deles é 11,2% no subátomo 4b.

Subátomo	Tempo médio de viagem (minutos)			Desvio em relação a simulação		Desvio em relação a amostra	
	Modelo	Simulação	Amostra	Minutos	%	Minutos	%
1a	8,900	8,744	9,000	0,156	1,8	-0,1000	-1,1
1b	8,900	8,890	8,500	0,010	0,1	0,4000	4,7
1c	8,700	8,892	8,500	-0,192	-2,2	0,2000	2,4
2a	10,700	9,753	10,300	0,947	9,7	0,4000	3,9
2b	10,800	10,939	11,500	-0,139	-1,3	-0,7000	-6,1
2c	10,400	10,946	11,100	-0,546	-5,0	-0,7000	-6,3
3a	8,900	10,532	9,000	-1,632	-15,5	-0,1000	-1,1
3b	9,000	8,894	8,700	0,106	1,2	0,3000	3,4
3c	9,600	8,894	9,500	0,706	7,9	0,1000	1,1
4a	10,700	11,239	11,000	-0,539	-4,8	-0,3000	-2,7
4b	10,900	10,607	9,800	0,293	2,8	1,1000	11,2
4c	11,000	10,607	10,100	0,393	3,7	0,9000	8,9
5a	9,500	9,822	9,000	-0,322	-3,3	0,5000	5,6
5b	10,100	9,500	9,200	0,600	6,3	0,9000	9,8
5c	9,600	9,499	9,300	0,101	1,1	0,3000	3,2
Média a	9,740	10,018	9,660	-0,278	-2,4	0,0800	0,9
Média b	9,940	9,766	9,540	0,174	1,8	0,4000	4,6
Média c	9,860	9,767	9,700	0,093	1,1	0,1600	1,8

Tabela 5.15 – Tempos médios de viagem para cada subátomo.

Os tempos médios de viagem para cada classe de usuário (átomos *a*, *b* e *c*) obtidos a partir do modelo, da simulação e da amostra estão na Tabela 5.16. Os desvios do modelo em relação à simulação e à amostra são baixos, todos menores que 5,0%.

Subátomo	Tempo Médio de viagem (minutos)		Desvio	
	Modelo	Simulação	Minutos	%
a	9,7100	9,6001	0,1099	1,1
b	9,9100	9,6393	0,2707	2,8
c	10,0100	9,8910	0,1190	1,2

Tabela 5.16 – Tempo médio de viagem nas classes *a*, *b* e *c*.

Os tempos médios de viagem de uma ambulância para os átomos *a*, *b* e *c* estão na Tabela 5.20. Não há diferenças nos tempos médios de viagem e de resposta para as ambulâncias nos átomos *a*, *b* e *c*. A comparação com a amostra não foi feita porque a quantidade de dados obtida não foi representativa, em alguns casos com

apenas uma ou nenhuma observação.

Ambulância	Tempo médio de viagem - Tarde			Tempo médio de resposta		
	<i>a</i>	<i>b</i>	<i>c</i>	<i>a</i>	<i>b</i>	<i>c</i>
1	9,6	9,8	9,9	9,6	9,8	9,9
2	8,7	8,8	8,9	8,7	8,8	8,9
3	8,7	8,9	9	8,7	8,9	9,0
4	11,6	11,3	11	11,6	11,3	11,0
5	11,6	10,9	11,6	11,6	10,9	11,6
6	9,6	8,9	9,4	9,6	8,9	9,4
7	8,9	9,4	9,7	8,9	9,4	9,7
8	9,9	12,8	9,4	9,9	12,8	9,4
9	9,2	9,3	9,3	9,2	9,3	9,3
10	9,3	9	9,2	9,3	9,0	9,2

Tabela 5.17 - Tempo médio de viagem e de resposta de cada ambulância nos átomos *a*, *b* e *c*.

(ii) Tempos de Resposta

A Tabela 5.18 mostra os resultados dos tempos médios de resposta para cada ambulância, obtidos pelo modelo e pela simulação. Os desvios do modelo em relação à simulação são baixos, todos menores que 10%.

Ambulância	Tempo médio de resposta (minutos)		Desvio em relação a simulação	
	Modelo	Simulação	Minutos	%
1	9,7100	9,5841	0,1259	1,3
2	8,8100	9,0651	-0,2551	-2,8
3	8,9100	9,0615	-0,1515	-1,7
4	11,2100	11,0691	0,1409	1,3
5	11,2100	11,1015	0,1085	1,0
6	9,1100	9,1983	-0,0883	-1,0
7	9,5100	9,1911	0,3189	3,5
8	11,3100	10,7487	0,5613	5,2
9	9,3100	9,2805	0,0295	0,3
10	9,1100	9,2745	-0,1645	-1,8
VSA	9,7100	9,5841	0,1259	1,3
VSB	9,8322	9,7767	0,0555	0,4
Total	9,8200	9,7574	0,0626	0,5

Tabela 5.18 – Tempos médios de resposta das ambulâncias.

A Tabela 5.19 mostra os resultados dos tempos médios de resposta para cada subátomo, obtidos pelo modelo e pela simulação. Os desvios do modelo em

relação à simulação são baixos, apenas nos subátomos *2a* e *3a* os desvios são maiores que 10%: 10,4% e 16,3%, respectivamente.

Subátomo	Tempo médio de resposta (minutos)		Desvio	
	Modelo	Simulação	Minutos	%
1a	8,9100	8,7493	0,1607	1,8
1b	8,9100	8,8941	0,0159	0,2
1c	8,7100	8,8992	-0,1892	-2,1
2a	10,7100	9,7579	0,9521	9,8
2b	10,8100	10,9437	-0,1337	-1,2
2c	10,4100	10,9530	-0,5430	-5,0
3a	8,9100	10,5367	-1,6267	-15,4
3b	9,0100	8,8983	0,1117	1,3
3c	9,6100	8,9010	0,7090	8,0
4a	10,7100	11,2435	-0,5335	-4,7
4b	10,9100	10,6113	0,2987	2,8
4c	11,0100	10,6140	0,3960	3,7
5a	9,5100	9,8269	-0,3169	-3,2
5b	10,1100	9,5043	0,6057	6,4
5c	9,6100	9,5058	0,1042	1,1
Média a	9,7500	10,0229	-0,2729	-2,4
Média b	9,9500	9,7703	0,1797	1,9
Média c	9,8700	9,7746	0,0954	1,1

Tabela 5.19– Tempos médios de resposta para cada subátomo.

Os tempos médios de resposta para cada classe de usuário (átomos *a*, *b* e *c*) obtidos a partir do modelo e da simulação estão na Tabela 5.19. Os desvios do modelo em relação à simulação são baixos, todos menores que 5,0%.

Subátomo	Tempo Médio de Resposta (minutos)		Desvio	
	Modelo	Simulação	Minutos	%
a	9,7100	9,6001	0,1099	1,1
b	9,9100	9,6393	0,2707	2,8
c	10,0100	9,8910	0,1190	1,2

Tabela 5.20 – Tempo médio de resposta nas classes *a*, *b* e *c*.

Comparando os desvios dos resultados das medidas de desempenho do modelo em relação aos resultados das medidas de desempenho da simulação, pode-se concluir que o modelo hipercubo é uma boa aproximação para o sistema, sem considerar os atendimentos de remoção de pacientes. Por outro lado, a comparação dos

resultados de tempos de espera e tempos de resposta do modelo e da amostra ficou prejudicada, devido à falta de informações mais precisas com relação aos apontamentos na amostra, e também devido à amostra considerar atendimentos de remoção de pacientes, o que distorceu para cima os tempos de espera dos usuários na amostra; consequentemente, também superestimou os tempos de resposta dos servidores na amostra.

As tabelas e as análises dos resultados para os períodos da manhã e noite são semelhantes com as do período da tarde e estão apresentados no Anexo E. Com base nesses resultados, pode-se observar que dependendo do período do dia (manhã, tarde e noite), há diferenças significativas que devem ser consideradas para a análise e configuração do SAMU-RP.

Resultados das medidas de desempenho para o período da manhã

A partir da análise das medidas de desempenho para o período da manhã, pode-se observar que os *workloads* desse período são maiores que no período da tarde. Observando as Tabelas 5.4 (taxa de chegada) e 5.9 (taxa de serviço), pode-se notar que apesar da taxa de chegada dos chamados ser maior no período da tarde, a taxa de serviço nesse período também é bem maior que no período da manhã. Dessa forma, o tempo de atendimento das ambulâncias no período da manhã é maior que no período da tarde, fazendo com que a taxa de utilização das ambulâncias seja maior no período da manhã. Pode-se notar que em uma análise clássica, quando apenas a taxa de chegada seria observada, o período da tarde seria o escolhido. A probabilidade de fila no sistema é muito baixa ($P_Q = 0,0038$) fazendo com que os tempos de viagem coincidam com os tempos de resposta no sistema.

Resultados das medidas de desempenho para o período da noite

A taxa de chegada dos chamados no período da noite é menor que nos períodos da manhã e tarde, e a taxa de serviço no período da noite é bem maior que nos outros períodos de forma que os chamados são atendidos mais rapidamente. Esse comportamento pode ser observado nas Tabelas 5.4 e 5.9. O período da noite é o menos congestionado dos três analisados, com probabilidade de fila menor que 10^{-4} de forma que os tempos de viagem e de resposta são coincidentes. O *workload* médio dos VSB's

é 0,19, enquanto que do VSA é 0,02. Apesar de ser o período menos congestionado, a análise desse período é importante e mostra um comportamento diferente dos períodos da manhã e da tarde, pois os *workloads* das ambulâncias localizadas fora da região central passam mais tempo ocupadas no período da noite. Os *workloads*, os tempos de viagem e os tempos de resposta são bem menores que nos outros períodos.

5.4.2 Cenário 1 – Atendimento dos chamados de remoção

Todos os VSB's do SAMU-RP podem fazer atendimentos de remoção (descritos na Seção 5.4). O objetivo desta análise é verificar o impacto dos atendimentos de remoção, avaliando como essa sobrecarga de trabalho nas ambulâncias afeta o desempenho do sistema. Para esta análise, foi considerado o cenário 1, uma vez que não foram coletados dados para comparação com a amostra. Na falta de dados disponíveis no SAMU-RP, este cenário considera algumas hipóteses simplificadoras: os chamados de remoção chegam aleatoriamente, de acordo com o processo de Poisson; os chamados são atendidos assim que uma ambulância fica desocupada, ou seja, sem política de agendamento; as remoções são classificadas como chamados de prioridade c ; os tempos de atendimento destes chamados são exponencialmente distribuídos. Na prática, o SAMU-RP realiza parte das transferências de pacientes de forma programada, levando em conta a disponibilidade das ambulâncias e, às vezes, removendo mais de um paciente na mesma viagem da ambulância.

A proporção de chamados do tipo remoção (1,6667 chamados/hora) foi acrescentada somente nos átomos c ($1c$, $2c$, $3c$, $4c$ e $5c$). Pode-se ver, na Tabela 5.21, que acrescentando atendimentos de remoção a taxa de chamados/hora, λ_j , aumentou consideravelmente a demanda no sistema, uma vez que, apesar da baixa prioridade, o volume desse tipo de atendimento é bastante representativo, da ordem de 60% de todos os atendimentos do sistema.

Além da taxa de chegada dos chamados, os tempos de atendimento de cada ambulância também podem mudar na presença de atendimentos de remoção. Porém, não foram coletados dados dos atendimentos de remoção, de forma que são utilizados os mesmos tempos de atendimento obtidos para o cenário original (Tabela 5.9). A Figura 5.11 mostra as taxas de utilização de cada ambulância (*workloads*), para o cenário 1. Os resultados mostram que as *workloads* são substancialmente altas (em torno de 80% para

os períodos da manhã e tarde e em torno de 70% para o período da noite), tratando-se de um serviço emergencial.

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

Tabela 5.21 – Taxas médias de chegada dos chamados para cada átomo para o cenário 1.

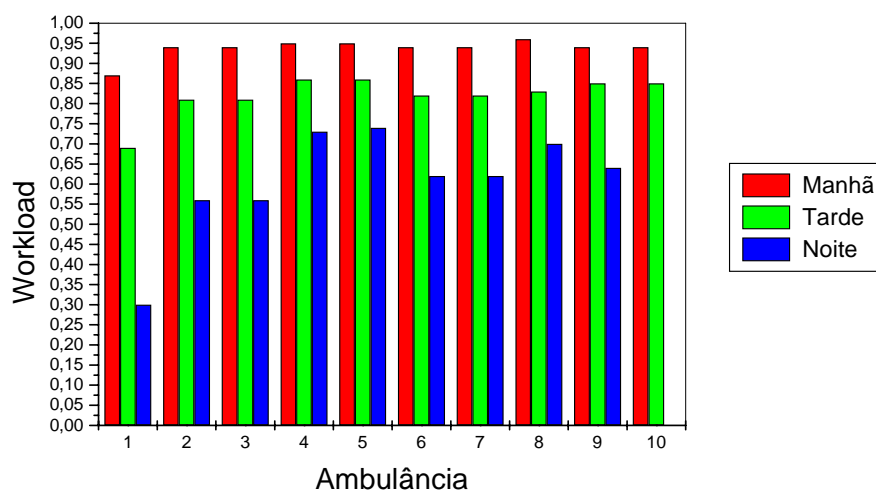


Figura 5.11 – *Workloads* das ambulâncias para o cenário 1.

No modelo de simulação, assim como no modelo hipercubo, foram consideradas distribuições exponenciais para o intervalo de tempo entre as chegadas e os serviços, utilizando os mesmos dados do modelo hipercubo. O objetivo da simulação neste cenário é validar os resultados obtidos pelo modelo hipercubo. Os tempos de resposta do cenário 1 não coincidem com os tempos de viagem. A Tabela 5.22 mostra o

tempo médio de espera em fila no sistema e para os átomos a , b e c para o período da tarde. Os desvios do modelo com a simulação são pequenos, todos menores que 5%. A probabilidade de todos os servidores estarem ocupados ($P_{1111111111}$) é de 0,1000 e a probabilidade de fila é de 0,3238.

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

Tabela 5.22 – Tempo médio de espera em fila para o cenário 1.

A Tabela 5.23 mostra os resultados do tempo médio de viagem e de resposta no sistema, além da comparação dos resultados do modelo hipercubo no cenário 1 e a simulação. O desvio do tempo médio de viagem no sistema do cenário 1 comparado com a simulação é de 0,4%, enquanto que o desvio do tempo médio de resposta no sistema do cenário 1 comparado com a simulação é de 0,6%.

		Tempo médio de viagem no sistema	Tempo médio de resposta no sistema
Modelo		10,1000	13,6340
Simulação		10,0476	13,5492
Desvio	Minutos	0,0524	0,0848
	%	0,5	0,6

Tabela 5.23 – Tempo médio de viagem e de resposta no sistema para o cenário 1.

A Tabela 5.24 mostra os *workloads* das ambulâncias do cenário 1 e da simulação, todos os desvios são menores que 5,0%.

Ambulância	Workload		Desvio	
	Modelo	Simulação	Minutos	%
1	0,6500	0,6523	-0,0023	-0,4
2	0,8000	0,8037	-0,0037	-0,5
3	0,8100	0,8024	0,0076	0,9
4	0,8400	0,8350	0,0050	0,6
5	0,8400	0,8337	0,0063	0,8
6	0,8100	0,8081	0,0019	0,2
7	0,8100	0,8067	0,0033	0,4
8	0,8200	0,8180	0,0020	0,2
9	0,8300	0,8334	-0,0034	-0,4
10	0,8300	0,8320	-0,0020	-0,2
VSA	0,6500	0,6523	-0,0023	-0,4
VSB	0,8211	0,8192	0,0019	0,2
Total	0,8040	0,8025	0,0015	0,2

Tabela 5.24 – Workload das ambulâncias do cenário 1.

A Tabela 5.25 mostra o tempo médio de viagem e de resposta para o cenário 1 e a comparação com a simulação. Todos os desvios foram menores que 10%.

Ambulância	Tempo médio de viagem (minutos)		Desvio		Tempo médio de resposta (minutos)		Desvio	
	Modelo	Simulação	Minutos	%	Modelo	Simulação	Minutos	%
1	10,1000	9,9534	0,1466	1,5	13,6340	13,4550	0,1790	1,3
2	9,8000	9,7494	0,0506	0,5	13,3340	13,2510	0,0830	0,6
3	9,8000	9,7386	0,0614	0,6	13,3340	13,2402	0,0938	0,7
4	9,9000	9,7044	0,1956	2,0	13,4340	13,2060	0,2280	1,7
5	10,2000	9,7200	0,4800	4,9	13,7340	13,2216	0,5124	3,9
6	10,4000	10,1946	0,2054	2,0	13,9340	13,6962	0,2378	1,7
7	10,0000	10,2060	-0,2060	-2,0	13,5340	13,7076	-0,1736	-1,3
8	10,8000	11,2608	-0,4608	-4,1	14,3340	14,7624	-0,4284	-2,9
9	9,8000	9,9762	-0,1762	-1,8	13,3340	13,4778	-0,1438	-1,1
10	10,0000	9,9516	0,0484	0,5	13,5340	13,4532	0,0808	0,6
VSA	10,1000	9,9534	0,1466	1,5	13,6340	13,4550	0,1790	1,3
VSb	10,0778	10,0557	0,0220	0,3	13,6118	13,5573	0,0544	0,4
Total	10,0800	10,0455	0,0345	0,4	13,6140	13,5471	0,0669	0,5

Tabela 5.25 – Tempos médios de viagem e de resposta das ambulâncias para o cenário 1.

Os tempos médios de viagem e de resposta do cenário 1 foram calculados para cada átomo, os resultados estão na Tabela 5.26. Os desvios do modelo em relação a simulação são menores que 10%.

Subátomo	Tempo médio de viagem (minutos)		Desvio		Tempo médio de resposta (minutos)		Desvio	
	Modelo	Simulação	Minutos	%	Modelo	Simulação	Minutos	%
1a	9,7000	9,3534	0,3466	3,7	11,1000	10,7388	0,3612	3,4
1b	9,7000	9,5652	0,1348	1,4	11,3000	11,1570	0,1430	1,3
1c	9,4000	9,5676	-0,1676	-1,8	13,6000	13,7238	-0,1238	-0,9
2a	9,7000	9,5388	0,1612	1,7	11,1000	10,9242	0,1758	1,6
2b	9,6000	9,7632	-0,1632	-1,7	11,2000	11,3550	-0,1550	-1,4
2c	9,7000	9,7620	-0,0620	-0,6	13,9000	13,9182	-0,0182	-0,1
3a	10,2000	10,3980	-0,1980	-1,9	11,6000	11,7834	-0,1834	-1,6
3b	10,1000	10,0056	0,0944	0,9	11,7000	11,5974	0,1026	0,9
3c	10,3000	10,0050	0,2950	2,9	14,5000	14,1612	0,3388	2,4
4a	11,5000	11,5278	-0,0278	-0,2	12,9000	12,9132	-0,0132	-0,1
4b	11,3000	11,4864	-0,1864	-1,6	12,9000	13,0782	-0,1782	-1,4
4c	11,4000	11,4834	-0,0834	-0,7	15,6000	15,6396	-0,0396	-0,3
5a	9,8000	9,9222	-0,1222	-1,2	11,2000	11,3076	-0,1076	-1,0
5b	9,8000	9,8646	-0,0646	-0,7	11,4000	11,4564	-0,0564	-0,5
5c	9,9000	9,8616	0,0384	0,4	14,1000	14,0178	0,0822	0,6
Média a	10,1800	10,1480	0,0320	0,4	11,5800	11,5334	0,0466	0,5
Média b	10,1000	10,1370	-0,0370	-0,3	11,7000	11,7288	-0,0288	-0,2
Média c	10,1400	10,1359	0,0041	0,0	14,3400	14,2921	0,0479	0,3

Tabela 5.26 – Tempos médios de viagem e de resposta para cada átomo para o cenário 1.

Os tempos médios de viagem e de resposta para os átomos *a*, *b* e *c* estão na Tabela 5.27. O maior desvio dos tempos médios de viagem e de resposta no cenário 1 comparado com a simulação é de 0,8%.

Subátomo	Tempo Médio de Viagem (minutos)		Desvio		Tempo Médio de Resposta (minutos)		Desvio	
	Modelo	Simulação	Minutos	%	Modelo	Simulação	Minutos	%
<i>a</i>	9,9000	9,8274	0,0726	0,7	11,3000	11,2128	0,0872	0,8
<i>b</i>	10,0000	9,8988	0,1012	1,0	11,6000	11,4906	0,1094	1,0
<i>c</i>	10,1000	10,1004	-0,0004	0,0	14,3000	14,2566	0,0434	0,3

Tabela 5.27 – Tempo médio de viagem e resposta para os átomos *a*, *b* e *c* para o cenário 1.

Os tempos médios de viagem e de resposta de cada ambulância para as classes *a*, *b* e *c* estão na Tabela 5.28.

Ambulância	Tempo médio de viagem - Tarde			Tempo médio de resposta		
	<i>a</i>	<i>b</i>	<i>c</i>	<i>a</i>	<i>b</i>	<i>c</i>
1	9,6	9,8	9,9	11,0	11,4	14,1
2	8,7	8,8	8,9	10,1	10,4	13,1
3	8,7	8,9	9	10,1	10,5	13,2
4	11,6	11,3	11	13,0	12,9	15,2
5	11,6	10,9	11,6	13,0	12,5	15,8
6	9,6	8,9	9,4	11,0	10,5	13,6
7	8,9	9,4	9,7	10,3	11,0	13,9
8	9,9	12,8	9,4	11,3	14,4	13,6
9	9,2	9,3	9,3	10,6	10,9	13,5
10	9,3	9	9,2	10,7	10,6	13,4

Tabela 5.28 - Tempo médio de viagem de cada ambulância para as classes *a*, *b* e *c* para o cenário 1.

Com essa análise, pôde-se observar que, em todas as medidas de desempenho analisadas para cenário 1, há um aumento considerável em relação aos tempos de viagem, de forma que o desempenho do sistema considerando atendimentos de remoção piora consideravelmente. Além disso, os tempos médios de resposta do modelo comparados com a amostra são bastante diferentes. Como comentado na Seção 2.1.2, os VSB's atendem os chamados de remoção, mesmo sendo com prioridade mais baixa do que os chamados "leves". Esse procedimento deve influenciar no desempenho das ambulâncias, uma vez que pode surgir um chamado de prioridade "*a*, *b* ou *c*" e a ambulância mais próxima pode estar ocupada atendendo uma remoção. Essa política de atendimento deve aumentar o tempo de espera na fila de todos os chamados. Como mencionado antes, durante o processo de coleta de dados, as remoções não foram

consideradas. Assim, a primeira análise (o modelo no cenário original), na verdade, é como um cenário "melhor" que o SAMU-RP.

No Anexo F, estão apresentadas as análises dos períodos da manhã e noite. Pode-se verificar que os *workloads* e os tempos de resposta do período da manhã são mais altos que no período da tarde. Caso a análise fosse feita com base em apenas um período, o período da tarde seria o escolhido. A análise do período da manhã ficaria então comprometida, já que é o período mais congestionado do dia. A taxa de chegada de chamados do período da manhã é menor que no período da tarde, ver Tabela 5.4. Porém, a taxa de serviço do período da manhã é bem menor que no período da tarde (ver Tabela 5.9). Pode-se notar que a utilização do sistema é dado por $\rho = \frac{\lambda}{m\mu}$, que depende da taxa de chegada e da taxa de serviço. Em entrevista realizada com o coordenador e a enfermeira chefe do SAMU, eles não souberam explicar essas diferenças nos tempos de atendimento dos chamados com relação ao período do dia. Uma possível justificativa seria as ambulâncias tenderem a atender mais rápido no período da tarde do que no período da manhã, devido ao fato de no período da tarde haverem, em média, mais chamados do que no período da manhã. É necessário um estudo mais detalhado para entender melhor esse comportamento dos tempos de atendimento das ambulâncias.

5.4.3 Cenário 2 – Aumento de demanda

No cenário 2, analisa-se o impacto do aumento de demanda nos períodos da manhã e tarde, resultando em maiores taxas de utilização do sistema. Os resultados das medidas de desempenho destes períodos estão na Seção 5.4.1. O objetivo é analisar o comportamento das medidas de desempenho do sistema sob dois pontos de vista: do gerente do sistema, para isso foi escolhida uma medida interna, o *workload*; e do usuário do sistema, na qual foram escolhidas duas medidas: o tempo médio de espera na fila; e o tempo médio de resposta em cada subátomo.

Aumento de demanda no período da manhã

A Tabela 5.29 mostra o aumento do *workload* com o aumento da demanda no período da manhã. A utilização média das ambulâncias passa de 46% (com aumento

de demanda de 10%) para 90% (com aumento de demanda de 150%).

Ambulância	Workload			
	10%	25%	50%	150%
1	0,1906	0,3377	0,3595	0,8250
2	0,4587	0,5325	0,6188	0,9020
3	0,4586	0,5317	0,6201	0,9015
4	0,5677	0,6294	0,7030	0,9264
5	0,5678	0,6341	0,7034	0,9271
6	0,4732	0,5511	0,6317	0,9090
7	0,4691	0,5520	0,6302	0,9091
8	0,5016	0,5868	0,6674	0,9257
9	0,4355	0,5496	0,6022	0,9020
10	0,4340	0,5480	0,6036	0,9012
VSA	0,1906	0,3377	0,3595	0,8250
VSB	0,4851	0,5684	0,6423	0,9116
Total	0,4557	0,5453	0,6140	0,9029

Tabela 5.29 – *Workloads* do cenário 2 - Aumento de demanda no período da manhã.

A Tabela 5.30 mostra os resultados dos tempos de espera na fila quando a demanda tem aumento de 10% até 150%. O tempo médio de espera na fila aumenta significativamente, especialmente na classe *c*, no período da manhã.

Tempo médio de espera em fila (min.)				
Demanda	Sistema	<i>a</i>	<i>b</i>	<i>c</i>
10%	0,1629	0,5391	1,1013	9,0667
25%	0,0989	0,3007	0,5519	3,1763
50%	0,1390	0,4591	0,8700	5,7846
150%	0,2219	0,7914	1,6575	16,1857

Tabela 5.30 – Tempos médios de espera na fila do cenário 2 - Aumento de demanda no período da manhã.

A Tabela 5.31 mostra os tempos médios de resposta para cada subátomo no período da manhã, com o aumento da demanda de 10% para 150%. Assim, como observado no tempo médio de espera na fila, os tempo médios de resposta nos átomos *c* aumentam significativamente, de 10,2 minutos (com aumento de demanda de 10%) para 26,3 minutos (com aumento de demanda de 150%).

	Tempo médio de resposta			
Subátomo	10%	25%	50%	150%
1a	8,8975	9,2375	9,5264	12,7585
1b	9,2513	9,6844	10,0731	15,4626
1c	9,3658	10,0324	10,9525	25,8651
2a	10,4256	10,5442	10,6372	12,8182
2b	10,8100	10,6049	11,2179	15,3891
2c	10,9569	11,0447	11,9329	25,8041
3a	9,8830	10,2063	10,4748	13,3770
3b	9,3464	10,0249	10,4532	15,9582
3c	9,4252	10,1243	11,1656	26,3411
4a	11,1415	11,2786	11,8796	14,6285
4b	11,2802	12,2252	13,0869	17,1840
4c	11,7137	11,2854	13,4764	27,6166
5a	9,6398	9,8325	10,2040	12,9122
5b	9,4246	9,9309	10,4967	15,5505
5c	9,4216	10,1948	11,1858	25,8619
Média a	9,9975	10,2198	10,5444	13,2989
Média b	10,0225	10,4941	11,0656	15,9089
Média c	10,1766	10,5363	11,7426	26,2978

Tabela 5.31 – Tempos médios de resposta para os subátomos do cenário 2 - Aumento de demanda no período da manhã.

Os resultados mostram que seria necessário aumentar a demanda em 150% de chamados das classes *a*, *b* e *c* para que as medidas de desempenho analisadas fiquem próximas das medidas de desempenho do cenário 1, que considerou o atendimento de remoção no sistema.

Aumento de demanda no período da tarde

A Tabela 5.32 mostra o aumento do *workload* com o aumento da demanda no período da tarde. A utilização média das ambulâncias passa de 38% (com aumento de demanda de 10%) para 81% (com aumento de demanda de 150%).

Ambulância	Workload			
	10%	25%	50%	150%
1	0,1410	0,1793	0,2630	0,7084
2	0,3354	0,3851	0,4756	0,7953
3	0,3396	0,3896	0,4737	0,8000
4	0,4828	0,5365	0,6150	0,8583
5	0,4837	0,5368	0,6155	0,8602
6	0,3315	0,3893	0,4811	0,7993
7	0,3316	0,3881	0,4779	0,8006
8	0,3328	0,3841	0,4746	0,8049
9	0,4973	0,5457	0,6230	0,8553
10	0,4980	0,5458	0,6224	0,8554
VSA	0,1410	0,1793	0,2630	0,7084
VSB	0,4036	0,4557	0,5399	0,8255
Total	0,3774	0,4280	0,5122	0,8138

Tabela 5.32 – *Workloads* do cenário 2 - Aumento de demanda no período da tarde.

A Tabela 5.33 mostra os resultados dos tempos de espera na fila quando a demanda tem aumento de 10% até 150% no período da tarde. O tempo médio de espera na fila aumenta significativamente, especialmente na classe *c*.

Tempo médio de espera em fila (min.)				
Demanda	Sistema	<i>a</i>	<i>b</i>	<i>c</i>
10%	0,0326	0,0812	0,2744	3,8676
25%	0,0213	0,0497	0,1511	1,4923
50%	0,0284	0,0692	0,2248	2,6515
150%	0,0424	0,1094	0,3911	6,5590

Tabela 5.33 – Tempos médios de espera na fila do cenário 2 - Aumento de demanda no período da tarde.

A Tabela 5.34 mostra os tempos médios de resposta para cada subátomo no período da tarde, com o aumento da demanda de 10% para 150%. Assim, como observado no tempo médio de espera na fila, o tempo médio de resposta nos átomos *c* aumentam significativamente, de 10,2 minutos (com aumento de demanda de 10%) para 26,3 minutos (com aumento de demanda de 150%).

Subátomo	Tempo médio de resposta			
	10%	25%	50%	150%
1a	8,7610	8,8131	8,9997	10,9390
1b	9,0453	9,1527	9,3341	12,2662
1c	8,9355	9,2028	9,4892	16,1068
2a	10,4567	10,3744	10,3960	11,0804
2b	10,9366	10,4885	10,3644	12,3364
2c	10,7441	10,9585	10,6505	16,3492
3a	9,6361	9,7460	9,9180	11,5884
3b	8,9608	9,0658	9,4002	12,5605
3c	9,1698	9,1268	9,8575	16,4626
4a	11,0615	11,1365	11,2769	12,9410
4b	11,1544	10,3954	10,5557	14,3074
4c	11,0821	11,3275	10,8059	18,2985
5a	9,5791	9,6922	9,7745	11,1936
5b	10,2883	10,3664	9,4743	12,3393
5c	9,1811	9,8897	10,0514	16,1908
Média a	9,8989	9,9524	10,0730	11,5485
Média b	10,0771	9,8938	9,8257	12,7620
Média c	9,8225	10,1011	10,1709	16,6816

Tabela 5.34 – Tempos médios de resposta para os subátomos do cenário 2 - Aumento de demanda no período da tarde.

Da mesma forma que no período da manhã, os resultados mostram que seria necessário aumentar a demanda em 150% de chamados das classes *a*, *b* e *c* para que as medidas de desempenho analisadas fiquem próximas das medidas de desempenho do cenário 1, que considerou o atendimento de remoção no sistema. O atendimento de remoções tem um papel socialmente importante para a cidade, o que sugere a análise de outros cenários.

Os próximos cenários consideram uma ambulância a menos em cada região para o atendimento das classes *a*, *b* e *c*. Pode-se notar que não é necessário que a ambulância retirada da análise seja desativada, porque ela pode ser realocada para fazer apenas atendimentos de remoção. O critério para a escolha da ambulância consiste na retirada de uma ambulância do átomo que apresenta a menor taxa de chamados, desde que a base contenha mais de uma ambulância do mesmo tipo. Vale salientar que o procedimento de calibração das taxas de serviço, descrito na Seção 5.3.9, foi aqui realizado para estes experimentos.

5.4.4 Cenário 3 – Período da manhã com uma ambulância a menos

O cenário 3 representa o período da manhã com uma ambulância a menos na região oeste. A localização das ambulâncias para este cenário está representada na Figura 5.12.

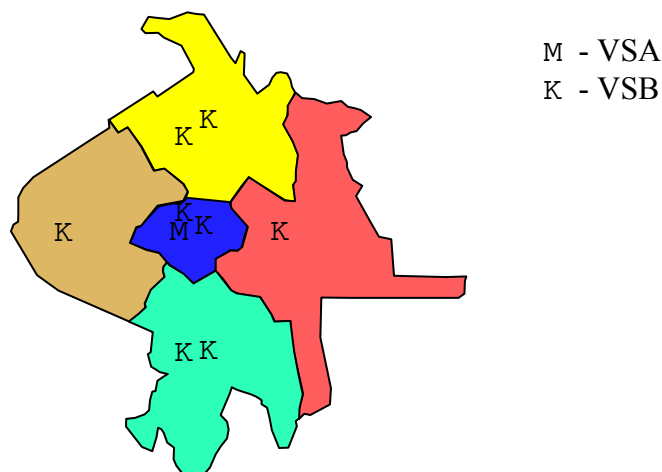


Figura 5.12 – Localização das ambulâncias no cenário 3.

A Tabela 5.35 contém os *workloads* das ambulâncias para o cenário 3 e a comparação com o SAMU-RP. Os resultados mostram que há aumento dos *workloads* das ambulâncias no cenário 3 em relação ao SAMU-RP, com desvios de, no mínimo 4,2% e, no máximo, 23,5%.

Ambulância	Workload		Desvio (%)
	cenário 3	SAMU-RP	
1	0,1945	0,1600	21,6
2	0,4548	0,4149	9,6
3	0,4761	0,4187	13,7
4	0,5570	0,5290	5,3
5	0,5500	0,5280	4,2
6	0,4646	0,4265	8,9
7	0,4705	0,4312	9,1
8	0,4997	0,4590	8,9
9	0,4868	0,3943	23,5
10	...	0,3902	...
VSA	0,1945	0,1600	21,6
VSB	0,4949	0,4435	10,4
Total	0,4616	0,4152	11,6

Tabela 5.35 - *Workloads* do cenário 3.

Os tempos médios de espera na fila para o cenário 3 e a comparação com o SAMU-RP estão na Tabela 5.36. Os resultados mostram que o aumento dos tempos médios de espera na fila no cenário 3 em relação ao SAMU-RP, em minutos, não são significativos com relação aos desvios absolutos. Porém, as porcentagens desses desvios são bastante altas de, no mínimo 66,5% e, no máximo, 69,3%.

		Tempo médio de espera em fila (min.)			
		Sistema	<i>a</i>	<i>b</i>	<i>c</i>
Cenário 3		0,2680	0,1613	0,2279	0,3667
SAMU-RP		0,0849	0,0541	0,0737	0,1124
Desvio	Minutos	0,1831	0,1072	0,1542	0,2543
	%	68,3	66,5	67,7	69,3

Tabela 5.36 – Tempos médios de espera na fila do cenário 3.

Os tempos médios de resposta nos subátomos para o cenário 3 e a comparação com o SAMU-RP estão na Tabela 5.37. Os resultados mostram que os desvios, em minutos e em porcentagem, do aumento dos tempos médios de espera na fila no cenário 3 em relação ao SAMU-RP não são significativos.

Tempo médio de resposta (minutos)				
Subátomo	Cenário 3	SAMU-RP	Desvio	
			Minutos	%
1a	8,9578	8,8394	0,1184	1,3
1b	9,2327	9,0951	0,1376	1,5
1c	9,3940	9,2070	0,1870	2,0
2a	10,4688	10,4075	0,0613	0,6
2b	11,0548	10,5856	0,4692	4,4
2c	11,1378	10,9857	0,1521	1,4
3a	9,9560	9,8009	0,1551	1,6
3b	9,3216	9,2072	0,1144	1,2
3c	9,4822	9,2090	0,2732	3,0
4a	11,1580	11,0539	0,1041	0,9
4b	11,3125	11,1987	0,1138	1,0
4c	10,5356	10,6731	-0,1375	-1,3
5a	9,8153	9,5742	0,2411	2,5
5b	9,8654	9,4547	0,4107	4,3
5c	10,6968	9,4739	1,2229	12,9
Média a	10,0712	9,9352	0,1360	1,4
Média b	10,1574	9,9083	0,2491	2,5
Média c	8,9578	8,8394	0,1184	1,3

Tabela 5.37 – Tempo médio de resposta para os subátomos do cenário 3.

Com a análise do cenário 3, de forma geral, a diminuição de uma ambulância na região oeste no período da manhã não tem grande impacto nas medidas de desempenho analisadas, principalmente considerando os desvios absolutos no tempo de resposta ao usuário.

5.4.5 Cenário 4 – Período da tarde com uma ambulância a menos

O cenário 4 representa o período da tarde com uma ambulância a menos na região sul. A localização das ambulâncias para este cenário está representada na Figura 5.13.

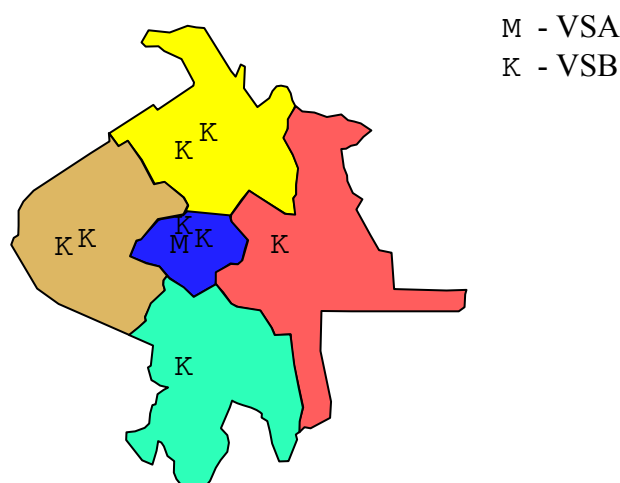


Figura 5.13 – Localização das ambulâncias no cenário 4.

A Tabela 5.38 contém os *workloads* das ambulâncias para o cenário 4 e a comparação com o SAMU-RP (cenário original). Os resultados mostram que há aumento dos *workloads* das ambulâncias no cenário 4 em relação ao SAMU-RP, com desvios de, no mínimo 5,9% e, no máximo, 34,7%.

Ambulância	Workload		Desvio (%)
	Cenário 4	SAMU-RP	
1	0,1464	0,1200	22,0
2	0,3599	0,3021	19,1
3	0,3464	0,3051	13,5
4	0,4793	0,4526	5,9
5	0,5031	0,4494	11,9
6	0,3503	0,2989	17,2
7	...	0,2977	...
8	0,4009	0,2831	34,7
9	0,3681	0,4637	30,0
10	0,5669	0,4637	22,3
VSA	0,1464	0,1200	22,0
VSB	0,4219	0,3773	19,3
Total	0,3913	0,3436	19,6

Tabela 5.38 – *Workloads* do cenário 4.

Os tempos médios de espera na fila para o cenário 4 e a comparação com o SAMU-RP estão na Tabela 5.39. Os resultados mostram que o aumento dos tempos médios de espera na fila no cenário 4 em relação ao SAMU-RP, em minutos, não são significativos com relação aos desvios absolutos. As porcentagens dos desvios são bastante altas de, no mínimo 77,5% e, no máximo, 78,7%.

		Tempo médio de espera em fila (min.)			
		Sistema	<i>a</i>	<i>b</i>	<i>c</i>
Cenário 4		0,0727	0,0466	0,0630	0,0956
SAMU-RP		0,0161	0,0110	0,0142	0,0204
Desvio	Minutos	0,0566	0,0356	0,0488	0,0752
	%	77,9	76,4	77,5	78,7

Tabela 5.39 – Tempos médios de espera na fila do cenário 4.

Os tempos médios de resposta nos subátomos para o cenário 4 e a comparação com o SAMU-RP estão na Tabela 5.40. Os resultados mostram que os desvios, em minutos, dos tempos médios de espera na fila no cenário 4 em relação ao SAMU-RP não são significativos com aumento de, no máximo, 3,3 minutos para chamados do subátomo 4*b*. As porcentagens dos desvios podem chegar a ser altas de, no máximo, 30,8% do subátomo 4*b*.

Tempo médio de resposta (minutos)				
Subátomo	Cenário 4	SAMU-RP	Desvio	
			Minutos	%
1a	8,7977	8,7412	0,0565	0,6
1b	8,9537	8,9509	0,0028	0,0
1c	9,0802	8,8740	0,2062	2,3
2a	10,4608	10,4541	0,0067	0,1
2b	10,9841	10,9359	0,0482	0,4
2c	10,7370	10,9951	-0,2581	-2,3
3a	10,9268	9,5894	1,3374	13,9
3b	10,8105	8,8027	2,0078	22,8
3c	10,9321	8,8028	2,1293	24,2
4a	12,6291	10,9483	1,6808	15,4
4b	13,8280	10,5697	3,2583	30,8
4c	13,4918	10,3793	3,1125	30,0
5a	9,7329	9,5792	0,1537	1,6
5b	9,3106	9,0212	0,2894	3,2
5c	9,8231	10,2192	-0,3961	-3,9
Média a	10,5095	9,8624	0,6470	6,3
Média b	10,7774	9,6561	1,1213	11,5
Média c	10,8128	9,8541	0,9588	10,1

Tabela 5.40 – Tempo médio de resposta do cenário 4.

Com a análise do cenário 4, de forma geral, a diminuição de uma ambulância na região oeste no período da manhã não tem grande impacto nas medidas de desempenho analisadas, principalmente considerando os desvios absolutos no tempo de resposta ao usuário, com exceção do subátomo 4b.

5.4.5 Cenário 5 – Período da noite com uma ambulância a menos

O cenário 5 representa o período da noite com uma ambulância a menos na região central. A localização das ambulâncias para este cenário está representada na Figura 5.14.

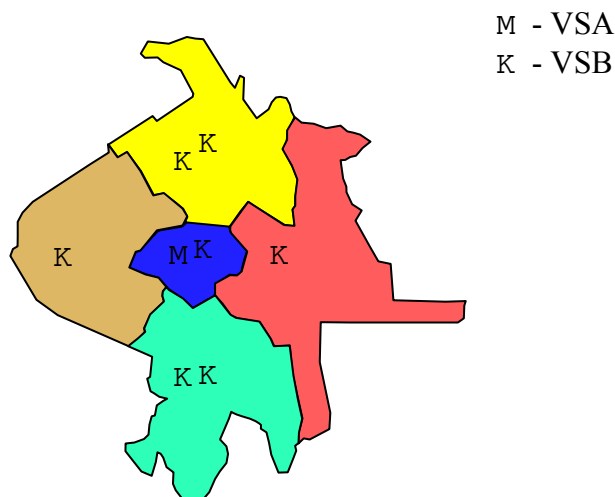


Figura 5.14 – Localização das ambulâncias no cenário 5.

A Tabela 5.41 contém os *workloads* das ambulâncias para o cenário 5 e a comparação com o SAMU-RP (cenário original). Os resultados mostram que há aumento dos *workloads* das ambulâncias no cenário 5 em relação ao SAMU-RP, com desvios de, no mínimo 9,5% e, no máximo, 61,04%.

Ambulância	Workload		Desvio (%)
	cenário 5	SAMU-RP	
1	0,0196	0,0179	9,5
2	0,1502	0,0941	59,6
3	...	0,0942	...
4	0,2364	0,2971	-20,4
5	0,3560	0,2972	19,8
6	0,2451	0,1522	61,0
7	0,1739	0,1505	15,5
8	0,2357	0,1962	20,1
9	0,2017	0,2296	-12,2
VSA	0,0196	0,0179	9,5
VSB	0,2036	0,1889	18,3
Total	0,1806	0,1699	17,2

Tabela 5.41 – *Workloads* do cenário 5.

Os tempos médios de espera na fila para o cenário 5 e a comparação com o SAMU-RP estão na Tabela 5.42. Os resultados mostram que o aumento dos tempos médios de espera na fila no cenário 5 em relação ao SAMU-RP, em minutos, não são significativos com relação aos desvios absolutos. Porém, as porcentagens dos desvios são bastante altas de, no mínimo 90,9% e, no máximo, 93,1%.

		Tempo médio de espera em fila (min.)			
		Sistema	<i>a</i>	<i>b</i>	<i>c</i>
cenário 5		0,0012	0,0010	0,0011	0,0014
SAMU-RP		0,0161	0,0110	0,0142	0,0204
Desvio	Minutos	-0,0149	-0,0100	-0,0131	-0,0190
	%	-92,5	-90,9	-92,3	-93,1

Tabela 5.42 – Tempos médios de espera na fila do cenário 5.

Os tempos médios de resposta nos subátomos para o cenário 5 e a comparação com o SAMU-RP estão na Tabela 5.43. Os resultados mostram que os desvios, em minutos, do aumento dos tempos médios de espera na fila no cenário 5 em relação ao SAMU-RP não são significativos com aumento de, no máximo, 4,6 minutos para chamados de origem no subátomo 4*b*. As porcentagens dos desvios podem chegar a ser altas de, no máximo, 46,3% no subátomo 4*b*.

Tempo médio de resposta (minutos)				
Subátomo	cenário 5	SAMU-RP	Desvio	
			Minutos	%
1a	8,9227	8,7004	0,2223	2,6
1b	9,1679	8,7201	0,4478	5,1
1c	9,1461	8,7321	0,4140	4,7
2a	10,0630	10,6225	-0,5595	-5,3
2b	10,2446	11,1133	-0,8687	-7,8
2c	10,5372	11,3280	-0,7908	-7,0
3a	10,9088	9,4952	1,4136	14,9
3b	10,8434	8,6532	2,1902	25,3
3c	10,8753	8,6522	2,2231	25,7
4a	12,8994	10,8571	2,0423	18,8
4b	14,5147	9,9218	4,5929	46,3
4c	14,2660	10,7870	3,4790	32,3
5a	9,8966	9,6312	0,2654	2,8
5b	10,4002	8,8880	1,5122	17,0
5c	10,4212	8,9048	1,5164	17,0
Média a	10,5381	9,8613	0,6768	6,7
Média b	11,0342	9,4593	1,5749	17,2
Média c	11,0492	9,6808	1,3683	14,5

Tabela 5.43 – Tempo médio de resposta do cenário 5.

Com a análise do cenário 5, de forma geral, a diminuição de uma ambulância na região oeste no período da manhã não tem grande impacto nas medidas de desempenho analisadas, principalmente considerando os desvios absolutos no tempo de resposta ao usuário, com exceção do subátomo 4b. Como mencionado antes, a ambulância retirada da análise pode ser realocada para fazer somente atendimentos de remoção, dado a importância social desse serviço.

A partir da análise do SAMU-RP por períodos, é possível encontrar configurações diferentes, como por exemplo, a quantidade e a localização das ambulâncias nesses períodos, a partir da análise das medidas de desempenho para cada um deles e avaliar suas características independentemente. Assim, é possível melhorar o desempenho do sistema, tanto do ponto de vista dos usuários, como do ponto de vista gerencial.

6 Conclusões

Esta tese propõe uma abordagem baseada na extensão do modelo hipercubo de filas espacialmente distribuídas para analisar a configuração do SAMU, utilizando atendimento com prioridade na fila e múltiplas configurações de localização de ambulâncias por dia. Em sistemas de atendimento emergencial mais congestionados, a fila é um fator importante a ser considerado. Esses sistemas diferenciam os usuários em classes e atribuem prioridades no atendimento de acordo com algum critério, por exemplo, a urgência do chamado do usuário, como no SAMU-RP. Para representar sistemas com essas características, neste estudo o modelo hipercubo foi estendido para considerar explicitamente políticas de prioridade no atendimento dos usuários. Foram desenvolvidas as equações de equilíbrio de cada estado do sistema e da cauda do modelo hipercubo, juntamente com suas medidas de desempenho, considerando prioridade na fila. A importância dessa extensão em relação ao modelo clássico foi ilustrada no final do Capítulo 4. Por meio dessa extensão, pôde-se obter medidas de desempenho por classe (por exemplo, os tempos de resposta aos usuários de cada classe), e não apenas medidas para todos os chamados agrupados, como em estudos anteriores com modelo hipercubo básico.

Um estudo de caso foi desenvolvido no SAMU-RP a partir de dados obtidos dos chamados ocorridos em 2005. Para aplicar a abordagem dessa tese, considerando vários períodos no dia de operação do SAMU, por exemplo, três períodos no dia: manhã, tarde e noite. Essa análise reforçou a importância de se considerar diferentes configurações do sistema para cada período do dia, e não apenas uma única configuração para todo o dia. Por meio dessa abordagem, pôde-se analisar as diferenças nas principais medidas de desempenho do sistema para cada período do dia. Por exemplo, caso a análise fosse feita com base apenas no período de pico de demanda do dia, o período escolhido seria a tarde. A partir dessa abordagem, foi observado que, apesar de o período da tarde ser o período com maior demanda, o período da manhã foi o mais desfavorável do ponto de vista de certas medidas de congestão do sistema, pois o tempo de atendimento dos chamados é maior nesse período.

Nesse estudo de caso, algumas hipóteses do modelo hipercubo não foram

validadas, por exemplo, os tempos de serviço das ambulâncias serem exponencialmente distribuídos. Foi então desenvolvido um modelo detalhado de simulação desse sistema que utilizou, entre outros, as distribuições dos tempos de serviço mais aderentes nos testes realizados no *software* Best-Fit®, com base em amostras do sistema. Os resultados do modelo foram então comparados com os resultados da simulação e da amostra, e os desvios foram, em geral, menores que 10%, o que validou o uso do modelo sob estas hipóteses. Os resultados mostraram ainda que os atendimentos de remoção não são independentes dos atendimentos das classes a , b e c nos horários de pico considerados. Como os atendimentos de remoção representam grande parcela dos atendimentos do SAMU-RP (cerca de 60%), é necessário um estudo mais detalhado desses chamados.

Além da configuração original do sistema SAMU-RP, foram também estudados diversos cenários para se avaliar alguns aspectos desse sistema. Por exemplo, o impacto do atendimento de remoção de pacientes no desempenho do sistema (cenário 1), a sensibilidade das medidas de desempenho com o aumento de demanda no período mais congestionado do dia (manhã) (cenário 2) e a análise de diferentes configurações de número e localização das ambulâncias no sistema para os períodos estudados (cenários 3, 4 e 5).

O cenário 1 foi desenvolvido para analisar o impacto do atendimento de remoção de pacientes no sistema. Para isso, as taxas de chegada dos átomos c foram aumentadas para avaliar o impacto de se atender remoções de pacientes. Convém salientar que isso ainda não é um cenário bem representativo dos atendimentos de remoção no SAMU-RP, uma vez que não foram coletados dados sobre os atendimentos de remoção, nem foram fornecidas informações suficientes para a estimativa das distribuições das chegadas e dos serviços desses atendimentos e, ainda, como eles influenciam os atendimentos das classes a , b e c . Além disso, as remoções podem ser agendadas e os atendimentos podem ser agrupados (por exemplo, uma ambulância pode fazer duas ou três remoções antes de voltar para a base), descaracterizando a aleatoriedade dos chamados e seus atendimentos, um a um (e não em lote), por uma ambulância. Assim, para fins de modelagem e análise desse cenário, foi considerado que os chamados de remoção chegam de acordo com a distribuição de Poisson e os tempos de serviço de acordo com a distribuição exponencial. Um modelo de simulação foi desenvolvido para validar os resultados dos modelos. As distribuições de chegada e

serviço utilizadas na simulação foram as mesmas do modelo hipercubo. Os resultados desse cenário indicam o quanto atendimentos de remoção podem piorar o desempenho do sistema. Por exemplo, os *workloads* passam de 42% para 90% no período da manhã e, no período da tarde, eles passam de 35% para 82% de utilização das ambulâncias. Os tempos médios de resposta para as classes *a*, *b* e *c* passam de 9,8min, 9,7min e 10,0min para 13,0min, 15,9min e 26,3min, respectivamente, no período da manhã e, no período da tarde, eles passam de 9,7min, 9,9min e 10,0min para 11,4min, 11,8min e 14,9min, respectivamente.

No cenário 2, o objetivo foi avaliar o impacto do aumento de demanda nos períodos da manhã e tarde, dos chamados das classes *a*, *b* e *c*, no período mais congestionado do dia (o período da manhã). Assim, a demanda dos chamados de todas as classes foi aumentada para 10%, 25%, 50% e 150%. Verificou-se que as medidas de desempenho são consistentes para aumentos de até 25% de demanda. Com aumentos de 50% ou mais, os tempos médios de resposta aumentaram relativamente, se comparados com os tempos médios de viagem. Quando a demanda dos chamados das classes *a*, *b* e *c* aumenta para 150%, as medidas de desempenho ficam próximas das medidas de desempenho obtidas a partir do cenário 1 (atendimento de remoções). Por exemplo, os *workloads* ficam em torno de 90% em ambos no período da manhã e 81% no período da tarde e, os tempos médios de resposta para as classes *a*, *b* e *c* chegam a 13,3min 15,9min e 26,3min no período da manhã e, no período da tarde, 11,5min 12,8min e 16,7min.

Os cenários 3, 4 e 5 avaliam a possibilidade de diferentes configurações para os períodos analisados. Optou-se pela retirada de uma ambulância de cada período, utilizando o seguinte critério: retirar uma ambulância da região com menor demanda, desde que ela tenha mais do que uma ambulância. Assim, o cenário 3 avalia o período da manhã com uma ambulância a menos na região oeste, o cenário 4 avalia o período da tarde com uma ambulância a menos na região sul e, por fim, o cenário 5 avalia o período da noite com uma ambulância a menos na região central. Os resultados obtidos foram comparados com o cenário original do SAMU-RP. Em todos os três cenários e em todas as medidas de desempenho, não foi observado aumento significativo dos tempos médios de viagem e resposta. Vale ressaltar que a ambulância desconsiderada dessa análise poderia ser destinada, por exemplo, para atender apenas chamados de remoção de pacientes, dada a importância social desses atendimentos.

6.1 Perspectivas de Pesquisa Futura

Durante o desenvolvimento dessa tese surgiram algumas perspectivas interessantes para pesquisas futuras. Algumas possibilidades seriam conduzir outras análises de sensibilidade além de alterações na demanda do sistema e estudar outros cenários alternativos explorando outras mudanças de configuração do SAMU-RP. Ainda com relação a alterações na demanda do sistema, também seria interessante estudar cenários aplicando técnicas de previsão para as taxas de chegada em cada átomo do sistema, por meio, por exemplo, de modelos de redes neurais artificiais, como em Setzler *et al.* (2008). No entanto, uma dificuldade para se aplicar essa abordagem seria a necessidade de amostras com muitos dados. Um aspecto favorável a estas pesquisas futuras é o fato de alguns sistemas SAMU no Brasil estarem começando a implementar sistemas GPS nas ambulâncias, o que irá simplificar o processo de coleta de dados de amostras e a precisão dessas informações.

Outra linha de pesquisa seria analisar o sistema SAMU considerando os atendimentos de remoção como uma quarta classe de usuários, uma vez que esses atendimentos podem representar uma parcela relevante do total de atendimentos do SAMU. Porém, um questionamento com relação a isso é se essas transferências deveriam mesmo ser realizadas pelo sistema SAMU, que é um sistema de atendimento eminentemente emergencial. Com esta abordagem, poder-se-á verificar se há diferenças entre os atendimentos de remoção. Por exemplo, as remoções entre hospitais podem ser mais importantes que as remoções de hospital para casa. Em muitas situações, a ambulância é redespachada para um atendimento sem que já tenha voltado para a base. Assim, uma pesquisa interessante seria também desenvolver uma abordagem do redespacho no modelo hipercubo para considerar essas situações. Como no SAMU-RP o VSA não atende a chamados b e c , também poderia-se estudar modificação no modelo hipercubo com prioridade na fila para considerar *backup* parcial e adaptar apropriadamente suas medidas de desempenho para um sistema com essas características.

Também seria interessante determinar o número mínimo de ambulâncias necessárias no sistema em diferentes períodos a fim de manter uma ou mais medidas de desempenho em um nível desejado. Por exemplo, obter o número mínimo de ambulâncias no sistema de forma que os tempos médios de resposta aos átomos a sejam

menores ou iguais a 12 minutos. Esta abordagem poderá ser útil para se eliminar os atendimentos de remoção de pacientes do sistema, as ambulâncias excedentes poderiam ser utilizadas para realizar apenas atendimentos de remoção. Uma alternativa para estas abordagens de otimização seria definir uma rede (grafo) com as possíveis localizações das bases e utilizar heurísticas de substituição de vértices, juntamente com o modelo hipercubo estendido para prioridade de fila, a fim de procurar pela melhor configuração para o sistema, com base em uma ou mais medidas de desempenho como, por exemplo, minimizando o tempo médio de resposta aos usuários e/ou maximizando o balanceamento dos *workloads* das ambulâncias.

Outra alternativa para essas análises seria investigar a possibilidade de incorporar o modelo hipercubo em problemas dinâmicos considerando a realocação e o reposicionamento das ambulâncias ao longo do dia. Por exemplo, pesquisar como incorporar o modelo hipercubo com prioridade de fila em problemas de realocação de máxima cobertura esperada, conforme em Gendreau *et al.* (2006) ou incorporá-lo em problemas de localização dinâmica de cobertura disponível, para determinar o número mínimo de ambulâncias e suas localizações para cada período do dia, em cada mudança significativa no padrão de demanda, conforme em Rajagopalan *et al.* (2008).

REFERÊNCIAS

- ALSALLOUMA O. I., RAND, G. K. (2006) Extensions to emergency vehicle location models. *Computers & Operations Research* 33, p. 2725–2743.
- ATKINSON J. B., KOVALENKO I. N., KUZNETSOV N. Yu., MIKHALEVICH K. V., (2006). Heuristic methods for the analysis of a queuing system describing emergency medical service deployed along a highway. *Cybernetics and Systems Analysis*, Vol. 42, No. 3, p. 379-391.
- ATKINSON J. B., KOVALENKO I. N., KUZNETSOV N. Yu., MIKHALEVICH K. V. (2008). A hypercube queueing loss model with customer-dependent service rates. *European Journal of Operational Research* 191, p. 223-239.
- BANKS, J. (1998) *Handbook of simulation*. John Wiley & Sons, Atlanta.
- BATTA R., DOLAN J.M., KRISHNAMURTHY N. M. (1989) The maximal expected covering location problem: Revised. *Transportation Science* 23, p. 277-287.
- BELL C., ALLEN D. (1969) Optimal planning of an emergency ambulance service. *Socio – Economic Planning Sciences* 3 (2), p. 95-101.
- BERMAN O., VANSUDEVA S. (2005) Approximating performance measures for public services. *Ieee Transactions on Systems, Man, and Cybernetcs*. Part A 35(4), p. 583-591.
- BODILY S. (1978) Police sector design incorporating preferences of interest groups for equality and efficiency. *Management Science* 24(12), p.1301-1313.
- BOFFEY B., GALVÃO R.D., ESPEJO L.G.A. (2007) A review of congestion models in the location of facilities with immobile servers. *European Journal of Operational Research*. 178, 643–662.
- BRANDEAU M., LARSON R.C. (1986) Extending and applying the hypercube queueing model to deploy ambulances in Boston. In: SWERSEY A.J, INGNALL E.J.(eds). *Delivery of Urban Services*. TIMS *Studies in the Management Science* 22, Elsevier, 121-153.
- BROTCORNE L., LAPORTE G., SEMET F. (2003). Ambulance location and relocation models. *European Journal of Operational Research* 147, p. 451-463.
- BURWELL T.H., JARVIS J.P., McKNEW M.A.(1993) Modeling co-located servers and dispatch ties in the hypercube model. *Computers & Operations Research* 20(2), 113-119.

- 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.
- CHAIKEN J., LARSON R. (1972) Methods for allocating urban emergency nits. *Management Science* 19(4), p.1280-1290.
- CHELST K. R. (1975). Implementing the hypercube queuing model in the new haven department of police services: a case study in technology transfer. *Rand Institute*. New York City, 91 p.
- CHELST K. R. (1978). An algorithm for deploying a crime directed patrol force. *Management Science* 24(12), p.1314-1327.
- CHELST K. R. (1981). Deployment of one-vs-two-officer patrol units. *Operations Research* 27(1), p. 199-204.
- CHELST K. R., 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. (2000) A statistical analisys of simulated annealing applied to the p-median problem. *Annals of Operational Research* 96, p. 61-74.
- CHIYOSHI F., GALVÃO R.D., MORABITO R. (2000) O uso do modelo hipercubo na solução de problemas de localização probabilísticos. *Gestão & Produção* 7(2), 146-174.
- 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 solution to the maximal expected covering location problem. *Computers and Operations Research* 30 (1), p. 87-96.
- CHURCH R. L., REVELLE C. (1974) The maximal covering location problem. *Papers Reg. Sci. Assoc.* 32, p. 101-118.
- CORRÊA F. A. (2008) Relaxações e método de depomposição para alguns problemas de localização de facilidades modelados em grafos. Instituto Nacional de Pesquisas Espaciais, INPE, de São José dos Campos. *Tese* (doutorado em Computação Aplicada).

- COSTA D.M. (2004) Uma metodologia iterativa para determinação de zonas de atendimento de serviços emergenciais. Universidade Federal de Santa Catarina. *Tese* (doutorado em Engenharia de Produção) - Departamento de Engenharia de Produção.
- COSTA NETO, P. L. O. (1977) *Estatística*. Edgard Blücher, 16ª reimpressão – 1998: São Paulo.
- DASKIN M. S. (1983). A maximum expected covering location model: Formulation, properties and heuristic solution. *Transportation Science* 17, p.48-70.
- DNIT, 2009 http://www.dnit.gov.br/menu/rodovias/estat_acid, acesso em 19 de setembro de 2009.
- EATON D. J., *et al.* (1985) Determining emergency medical service vehicle deployment in Austin, Texas. *Interfaces* 15(1), p. 96-108.
- ERKUT E., INGOLFSSON A., ERDOGAN G. (2008) Ambulance location for maximal survival. *NavalResearch Logistics* 55, pp. 42-58.
- FIGUEIREDO A. P. S., LORENA L. A. N. (2005) Localização de ambulâncias: uma aplicação para a cidade de São José dos Campos. *Anais XII Simpósio Brasileiro de Sensoriamento Remoto*, Goiânia, Brasil, INPE, p. 1965-1972.
- 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, p. 15-33.
- GALVÃO R. D., MORABITO, R. (2008). Emergency service systems: The use of the hypercube queuing model in the solution of probabilistic location problems. *International Transactions in Operational Research* 15, p. 1-25.
- GENDREAU M., LAPORTE G., SEMET, F. (2006) The maximal expected coverage relocation problem for emergency vehicles. *Journal of the Operational Research Society* 57, p. 22–28.

- GEROLIMINIS N., KARLAFTIS M. G., SKABARDONIS A. A generalized queuing hypercube model for locating emergency response vehicles in urban transportation networks. *TRB 2006 Annual Meeting*.
- GOLDBERG J., DIETRICH R., CHEN J. M., 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, p. 125-141.
- 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), p. 964-981.
- HAKIMI S.L. 1964, Optimum locations of switching centers and the absolute centers and medians of a graph, *Operations Research* 12, pp. 450-459.
- IANNONI A. P. (2005) Otimização da configuração e operação de sistemas médico emergenciais em rodovias utilizando o modelo hipercubo. Universidade Federal de São Carlos. *Tese* (doutorado em Engenharia de Produção) - Departamento de Engenharia de Produção.
- IANNONI A. P., MORABITO R. (2006) Modelo hipercubo integrado a um algoritmo genético para análise de sistemas médicos emergenciais em rodovias. *Gestão & Produção*, 13(1), 93-104.
- IANNONI A. P., MORABITO R. (2006) Modelo de fila hipercubo com múltiplo despacho e *backup* parcial para análise de sistemas de atendimento médico emergenciais em rodovias. *Pesquisa Operacional*, v.26, n.3, p.493-519.
- IANNONI, A. P., MORABITO, R., SAYDAM, C. (2008a) A hypercube queueing model embedded into a genetic algorithm for ambulance deployment on highways. *Annals of Operations Research* 157 (1), p. 207 – 224.

- IANNONI A. P., MORABITO, R., SAYDAM, C. (2008b) An optimization approach for ambulance location and the districting of the response segments on highways. *European Journal of Operational Research* doi:10.1016/j.ejor.2008.02.003. 27
- IANNONI, A. P., MORABITO, R. (2008) Otimização da localização das bases de ambulâncias e do dimensionamento das suas regiões de cobertura em rodovias. *Produção Online*. v. 18, n. 1, pp. 47-63.
- IBGE, 2008, Diretoria de Pesquisas, Coordenação de População e Indicadores Sociais, Pesquisa de Informações Básicas Municipais.
- 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 dispatch of fire companies. *Management Science*. 28(4), p.366-378.
- INGOLLFSSON A., BUDGE S, ERKUT E. (2008) Optimal ambulance location with random delays and travel times. *Health Care Management Science*, forthcoming.
- JARVIS J. P. (1985). Approximating the equilibrium behavior of multi-server loss systems. *Management Science*. 31, p.235-239.
- JOHNSON, N. L., KOTZ, S., and, BALAKRISHNAN, N. (1994). *Univariate continuous distributions*. Vol. 1, 2nd ed. New York: John Wiley & Sons.
- JOHNSON, N. L., KOTZ, S., and, BALAKRISHNAN, N. (1995). *Univariate continuous distributions*. Vol. 2, 2nd ed. New York: John Wiley & Sons.
- KENDALL, D. G. (1953) Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded markov chain. *The Annals of Mathematical Statistics* 24(3), 338-354.
- KLEINROCK L. (1975) *Queuing Systems*. Volume I: Theory. New York: John Wiley & Sons.
- KLEINROCK L. (1976) *Queuing Systems*. Volume II: Computer Applications. New York: John Wiley & Sons.
- KOLESAR P., BLUM E. (1973) Square root laws for the 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. J., IGNALL E. TIMS Studies in the Management Science 22, p. 87-119. Elsevier. North-Holland.
- LARSON R.C. (1971) Measuring the response patterns of New York city police patrol cars. *Rand Institute*. New York City, 73 p.
- LARSON R.C. (1974) Hypercube queuing model for facility location and redistricting in urban emergency services. *Computers and operations research* 1, 67-95.
- LARSON R.C. (1975) Approximating the performance of urban emergency service systems. *Operations Research* 23, 845-868.
- LARSON R.C., ODoni A.R. (2007) *Urban Operations Research*. 2 ed. Dynamic Ideas, Belmont, Massachusetts.
- LARSON R., MCKNEW M.A. (1982) Police patrol-initiated activities within a systems queuing model. *Management Science* 28(7), 759-774.
- LOPES S. L. B., FERNANDES R. J. (1999). Uma breve revisão do atendimento médico pré-hospitalar. Simpósio: *Trauma II*, Medicina, Ribeirão Preto 32, p. 381-387.
- LOUVEAUX F. (1993). Stochastic location analysis. *Location Science* 1, p. 127-154.
- LUQUE L. (2006). Análise da aglutinação de estados em cadeias de markov do modelo hipercubo de filas com servidores co-localizados. Dissertação de Mestrado. INPE – São José dos Campos.
- MAGALHÃES, M. N., LIMA, A. C. P. (2002) *Noções de probabilidade e estatística*. Edusp, 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.
- MARIANOV V., SERRA D. (1998). Probabilistic maximal covering location-allocation for congested system. *Journal of Regional Science* 38, 401-424.
- MARIANOV V., RÍOS M., (2000). A probabilistic quality of service constraint for a location model of switches in ATM communications networks. *Annals of Operations Research* 96, 237-243.

- MARIANOV V., SERRA D. (2001a) Hierarchical location-allocation models for congested systems. *European Journal of Operational Research* 135, 195–208.
- MARIANOV V., SERRA D. (2001b) Location models for airline hubs behaving as M/D/c queues. *Computers and Operations Research* 30, 983–1003.
- MARIANOV V., SERRA D. (2003) Location-allocation of multiple-server service centers with constrained queues or waiting times. *Annals of Operations Research* 111, 35–50.
- MENDONÇA F. (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., MORABITO R. (2000) Aplicação do modelo hipercubo para análise de um sistema médico-emergencial em rodovia, *Gestão & Produção* 7(1), 73-91.
- MENDONÇA F.C., MORABITO R. (2001) Analyzing emergency service ambulance deployment on a Brazilian highway using the hypercube model. *Journal of the Operation Research Society* 52, 261-268.
- MINISTÉRIO DA SAÚDE (2009) - http://portal.saude.gov.br/portal/saude/visualizar_texto.cfm?idtxt=23745&janela=1, (data de acesso: 02/10/2009)
- OLIVEIRA L. K. (2003) Uma aplicação do modelo hipercubo de filas para avaliação do centro de emergência da polícia militar de Santa Catarina. Florianópolis. *Dissertação* (Mestrado em Engenharia de Produção) Departamento de Engenharia de Produção, Universidade Federal de Santa Catarina.
- OMS, 2009, <http://www.portaldotransito.com.br/noticias/brasil-e-quinto-pais-do-mundo-em-mortes-por-acidentes-de-transito.html>. Acesso em 19 de setembro de 2009.
- OPAS, 2008, http://www.opas.org.br/sistema/arquivos/carta_lei_seca.pdf, acesso em 19 de setembro de 2009.
- OWEN S., DASKIN M. S. (1998). Strategic facility location: a review. *European Journal of Operational Research* 3(2), p. 423-447.
- PEGDEN, C. D. SHANNON, R. E. SADOSWSKI, R. P. (1995) *Introduction to Simulation Using SIMAN*. 2.ed.McGraw-Hill, New York.

PORTARIA Nº 1863/GM Em 29 de setembro de 2003.
<http://dtr2001.saude.gov.br/sas/PORTARIAS/Port2003/GM/GM-1863.htm>. Acesso em 22 de setembro de 2009.

PORTARIA Nº 1864/GM. Em 29 de setembro de 2003,
<http://dtr2001.saude.gov.br/sas/PORTARIAS/Port2003/GM/GM-1864.htm>. Acesso em 22 de setembro de 2009.

PREFEITURA DE CAMPINAS (2008) –

http://www.campinas.sp.gov.br/saude/unidades/samu/samu_cap_inst.htm, (data de acesso: 21/10/08).

RAJAGOPALAN, H. K. SAYDAM, C. XIAO, J. (2008) A multiperiod set covering location model for dynamic redeployment of ambulances. *Computers & Operations Research* 35, p. 814 – 826.

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.

RIDER K. (1976). A paramedic model for the allocation of fire companies in New York City. *Management Science* 23(2), p. 146-158.

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

SAVAS E. (1969). Simulation and cost-effectiveness analysis of New York's emergency ambulance service. *Management Science* 15(12), B608-B627.

SAYDAM C. AYTUG H. (2003) Accurate estimation of expected coverage: revised. *Socio-Economic Planning Sciences* 37, 69-80.

SAYDAM C., REPEDE J., BURWELL T. (1994) Accurate estimation of expected coverage: a comparative study. *Socio-Economic Planning Sciences* 28 (2), 113-120.

SAVAS E. (1969) Simulation and cost-effectiveness analysis os New york's emergency ambulance service. *Management Science* 15(12), B608-B627.

SETZLER H., SAYDAM, C., SUNGJUNE P. (2008) EMS Call Volume Predictions: A Comparative Study. *Computers & Operations Research* 36, p. 1843 – 1851.

- SIMPSON N.C., HANCOCK P.G. (2009) Fifty years of operational research and emergency response. *Journal of the Operational Research Society* 60, p. 126-139.
- SINGER M., DONOSO P. (2008) Assessing an ambulance service with queuing theory. *Computers & Operations Research* 35, p. 2549 – 2560.
- 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, 151-200.
- TAKEDA R. (2000) Uma contribuição para avaliar o desempenho de sistemas de transporte emergencial de saúde. Universidade de São Paulo. *Tese* (doutorado em Transportes) – Escola de Engenharia de São Carlos.
- TAKEDA R.A., WIDMER, J.A., MORABITO, R. (2004) Aplicação do modelo hipercubo de filas para avaliar a descentralização de ambulâncias em um sistema urbano de atendimento médico de urgência. *Pesquisa Operacional* 24 (1), 39-72.
- TAKEDA R.A., WIDMER, J.A., MORABITO, R. (2007) Analysis of ambulance decentralization in an urban emergency medical service using the hypercube queueing model. *Computers & Operations Research* 34, p. 727-741.
- 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.
- TEIZ M. BART P. (1968) Heuristic methods for estimating the generalized vertex median of a weighted graph. *Operations Research* 16(5), p. 901-1092.
- WOLFF, R. W. (1989) *Stochastic modeling and theory of queues*. New Jersey, Prentice-Hall.
- ZAKI, A. S. CHENG, H. K. (1997) A simulation model for the Analysis and Management of an Emergency Service System. *Socio-Economics Planning Sciences* 31, p. 173-189.

ANEXO B

A Análise de Variância: ANOVA – Um Fator.

Considere A um fator, tal que A tenha k níveis (ou tratamentos) fixo (as conclusões não podem ser estendidas para níveis não considerados no experimento). O objetivo é avaliar se existe diferença significativa entre as médias dos níveis do fator. A partir de todas as combinações dos níveis, pode-se obter uma apresentação conforme a tabela B1 a seguir:

Nível	Fator A				Somas	Médias
1	y_{11}	y_{12}	...	y_{1n_1}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	...	y_{2n_2}	$y_{2.}$	$\bar{y}_{2.}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
k	y_{k1}	y_{k2}	...	y_{kn_k}	$y_{k.}$	$\bar{y}_{k.}$

Tabela B1: Apresentação dos dados para um fator.

O modelo que corresponde à Análise de Variância de um único fator é dado por:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \text{ onde:}$$

$$j = 1, \dots, n_i \text{ e } i = 1, 2, \dots, k;$$

μ é a média geral dos dados;

α_i é o efeito no nível i de A;

ε_{ij} é a componente aleatória do erro.

Seja $y_{i.}$ a soma das observações do i -ésimo nível de A, $\bar{y}_{i.}$ a média das observações do nível i de A, $y_{..}$ a soma de todas as observações e $\bar{y}_{..}$ a média geral das observações.

Expresso matematicamente,

$$y_{i.} = \sum_{j=1}^{n_i} y_{ij}, \quad \bar{y}_{i.} = \frac{y_{i.}}{n_i} \text{ e } i = 1, 2, \dots, k.$$

$$y_{..} = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}, \quad \bar{y}_{..} = \frac{y_{..}}{N}, \text{ onde } N \text{ é o número total de observações.}$$

Como no experimento tem-se k níveis do fator A, com $k - 1$ graus de liberdade, temos que a soma dos quadrados relativo aos níveis é:

$$SQ_N = \sum_{i=1}^k (\bar{y}_{i.} - \bar{y}_{..})^2.$$

Dentro de cada nível, temos n_i réplicas fornecendo $n_i - 1$ graus de liberdade para cada estimativa da variabilidade devido ao erro experimental, $\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$. Assim, a soma de quadrados devido ao erro experimental é:

$$SQ_E = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2. \text{ Ainda, } SQ_T = SQ_N + SQ_E.$$

Mais detalhes deste modelo podem ser vistos em COSTA NETO (1998) e MAGALHÃES & LIMA (2002). A tabela 2 resume a Análise de Variância. Um teste é realizado para verificar a igualdade dos efeitos dos k níveis. As hipóteses adequadas são:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k,$$

H_1 : pelo menos uma das médias diferente.

Fonte de variação	Soma de Quadrados	Graus de liberdade	Quadrado médio	F	F_α
Entre amostras	SQ_N	$k - 1$	$MQ_N = \frac{SQ_N}{k-1}$	$F = \frac{MQ_N}{MQ_E}$	$F_{k-1, k(n-1), \alpha}$
Residual	SQ_E	$k(n - 1)$	$MQ_E = \frac{SQ_E}{k(n-1)}$		
Total	SQ_T	$nk - 1$			

Tabela B2: Análise de Variância para um fator.

Assim, a Análise de Variância para um fator pode ser considerado um teste F com nível de significância α (normalmente, $\alpha = 0,05$). Outra forma de análise do teste é pelo P-valor, ou seja, a probabilidade de que a estatística do teste (como variável aleatória) tenha valor extremo em relação ao valor observado desta estatística, quando a hipótese H_0 é verdadeira. O P-valor é dado pela equação:

$P\text{-valor} = P[F(k-1, k(n-1)) > F_0 | H_0]$. Podemos dizer que o P-valor é o menor nível de confiança (α) para o qual rejeitamos a hipótese H_0 .

A análise estatística para um problema de k médias termina se não rejeitamos H_0 . Caso rejeitemos H_0 , há evidência de que pelo menos dois níveis em estudo diferem

significativamente. Dessa forma, é importante continuar a análise para identificar as diferenças entre as médias dos níveis por meio das comparações múltiplas. O procedimento mais eficiente parece ser o proposto por Tukey (COSTA NETO, 1998), mostrado a seguir.

Teste de Tukey (TSD – *Tukey Significant Difference*)

O Teste de Tukey permite fazer comparações múltiplas, sempre entre duas médias dos níveis. A estatística do teste pode ser definida por:

$$TSD = q_{\alpha}(k, n - K) \sqrt{\frac{QME}{n}}, \text{ para dados balanceados; e,}$$

$$TSD = \frac{q_{\alpha}(k, n - K)}{\sqrt{2}} \sqrt{QME \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}, \text{ para dados não balanceados.}$$

onde q é chamada de amplitude “studentizada” (pode ser obtida a partir de tabela elaboradas para o teste), que depende do número de níveis (k) e do número de graus de liberdade dos erros ($N - k$). O teste preserva o nível de significância para todos os contrastes. Rejeita-se a igualdade entre dois níveis se:

$$|\bar{y}_i - y_j| > TSD.$$

Aplicação da ANOVA para as observações ao longo do ano de 2005.

Os testes da Análise de Variância foram realizados pelo *software* Minitab^(R) 14. Os resultados podem ser vistos na tabela B3.

Análise de Variância					
Fonte de variação	Soma de Quadrados	Graus de liberdade	Quadrado médio	F	P-valor
Meses	11	11013	1001	3,94	0,000
Erro	353	89777	254		
Total	364	100790			

Tabela B3: Análise de Variância para as médias dos meses de 2005.

A partir da análise da tabela 3, com P-valor próximo de zero, rejeitamos a hipótese H_0 para igualdade das médias dos meses. Assim, foi aplicado o Teste de Tukey, de comparações múltiplas (tabela 4), para verificar quais médias são diferentes. Podemos

notar que apenas a média do mês de abril pode ser considerada diferente das médias dos demais meses.

	Abr	Ago	Dez	Fev	Jan	Jul	Jun	Mai	Mar	Nov	Out
Ago	0,95										
	27,63										
Dez	5,85	-8,33									
	32,54	18,14									
Fev	-1,94	-16,12	-21,03								
	25,44	11,04	6,14								
Jan	1,75	-12,43	-17,33	-10,24							
	28,44	14,04	9,14	16,93							
Jul	4,79	-9,39	-14,3	-7,2	-10,2						
	31,47	17,07	12,17	19,96	16,27						
Jun	0,81	-13,37	-18,27	-11,17	-14,17	-17,21					
	27,72	13,32	8,42	16,21	12,51	9,48					
Mai	2,14	-12,04	-16,94	-9,85	-12,85	-15,88	-12,13				
	28,83	14,43	9,52	17,32	13,62	10,59	14,56				
Mar	3,75	-10,43	-15,33	-8,24	-11,23	-14,27	-10,51	-11,62			
	30,44	16,04	11,14	18,93	15,23	12,2	16,17	14,85			
Nov	10,01	-4,17	-9,07	-1,97	-4,97	-8,01	-4,25	-5,36	-6,97		
	36,92	22,52	17,62	25,41	21,71	18,68	22,65	21,33	19,71		
Out	1,11	-13,07	-17,97	-10,88	-13,88	-16,91	-13,16	-14,27	-15,88	-22,36	
	27,79	13,39	8,49	16,28	12,59	9,56	13,53	12,2	10,59	4,33	
Set	6,78	-7,4	-12,3	-5,21	-8,21	-11,24	-7,49	-8,59	-10,21	-16,69	-7,56
	33,69	19,29	14,38	22,17	18,48	15,45	19,42	18,09	16,48	10,22	19,12

Tabela B4: Teste de Tukey, de comparações múltiplas, para os números médios de atendimentos dos meses de 2005.

A partir da análise da tabela B5, com P-valor igual a 0,86, não rejeitamos a hipótese H_0 para igualdade das médias dos atendimentos dos dez dias observados em agosto de 2005.

Análise de Variância					
Fonte de variação	Soma de Quadrados	Graus de liberdade	Quadrado médio	F	P-valor
Dias	9	39,85	4,43	0,52	0,860
Erro	230	1964,00	8,54		
Total	239	2003,85			

Tabela B5: Análise de Variância para os números médios de atendimentos dez dias observados em agosto de 2005.

ANEXO C

As figuras C1 a C12, a seguir, mostram o número de chamados (frequência) ao longo dos meses de 2005.

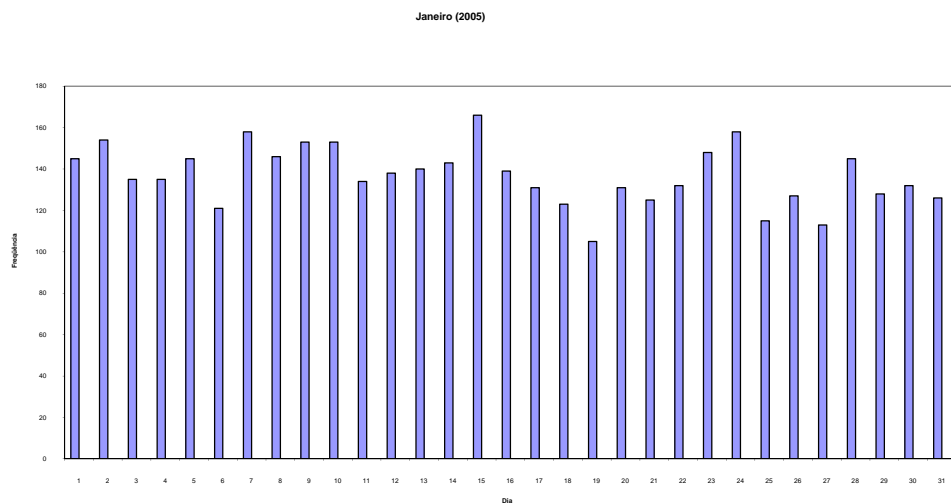


Figura C1 – Atendimentos de urgência e emergência realizados pelo SAMU_RP em janeiro de 2005.

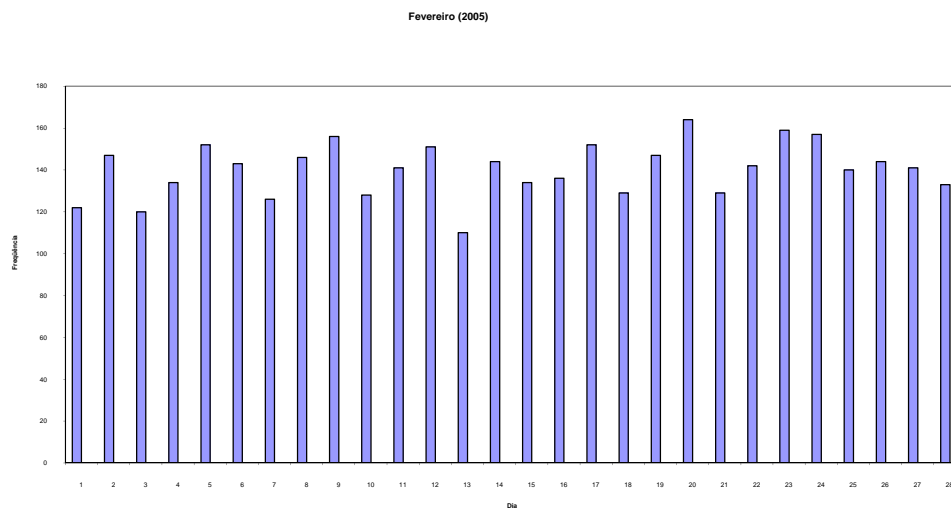


Figura C2 – Atendimentos de urgência e emergência realizados pelo SAMU_RP em fevereiro de 2005.

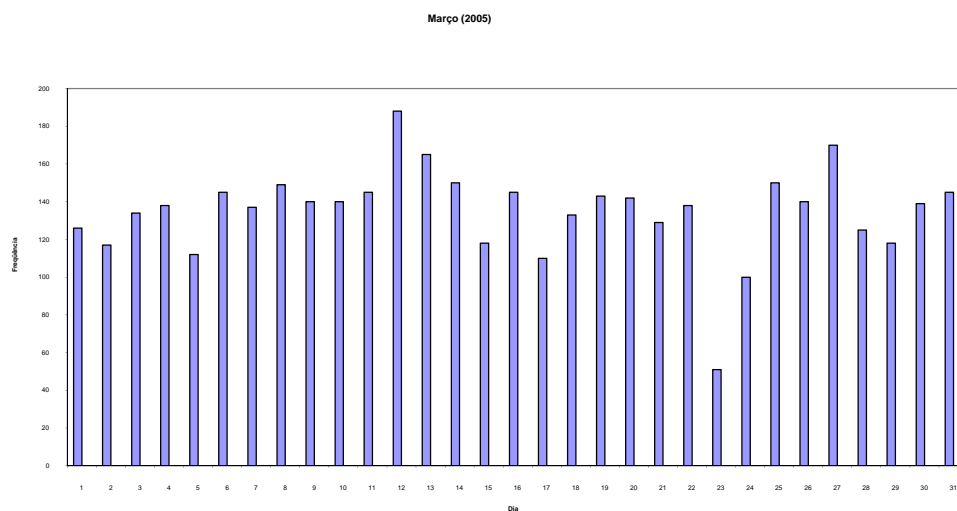


Figura C3 – Atendimentos de urgência e emergência realizados pelo SAMU_RP em março de 2005.

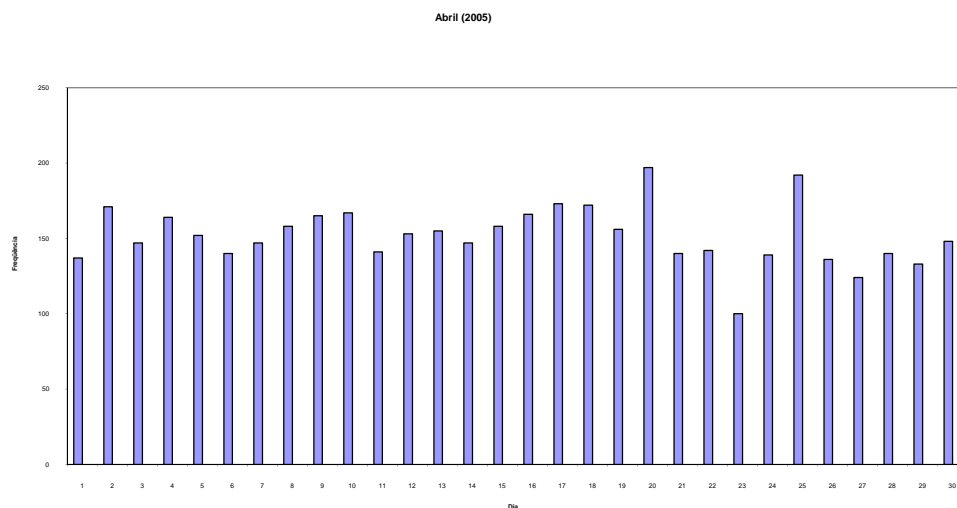


Figura C4 – Atendimentos de urgência e emergência realizados pelo SAMU_RP em abril de 2005.

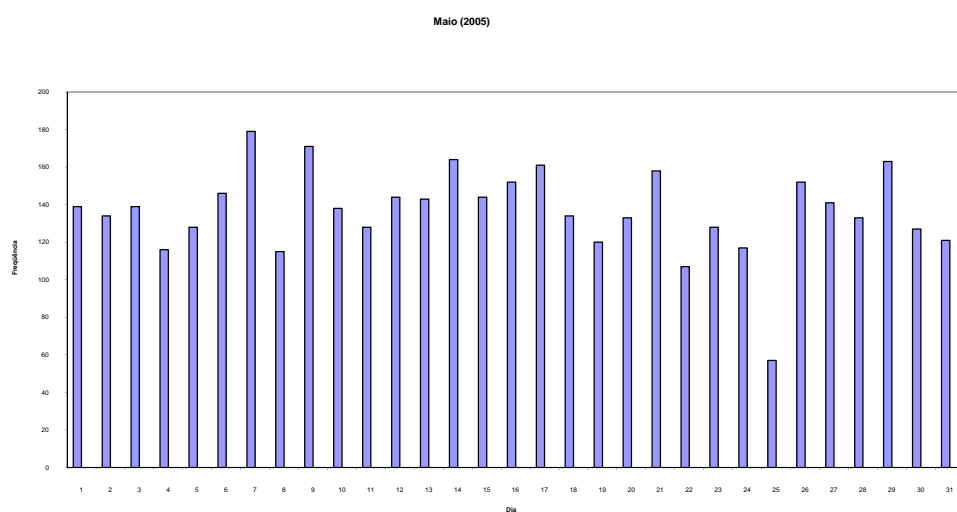


Figura C5 – Atendimentos de urgência e emergência realizados pelo SAMU_RP em maio de 2005.

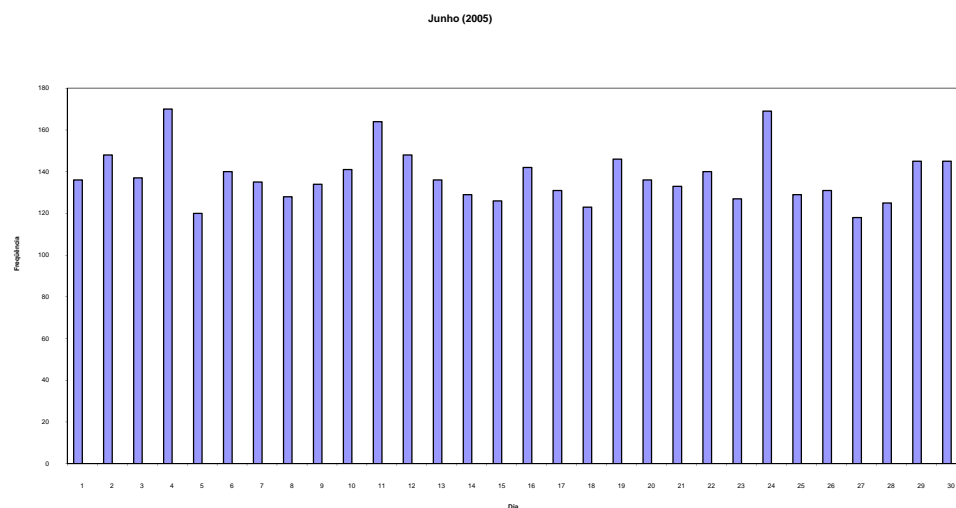


Figura C6 – Atendimentos de urgência e emergência realizados pelo SAMU_RP em junho de 2005.

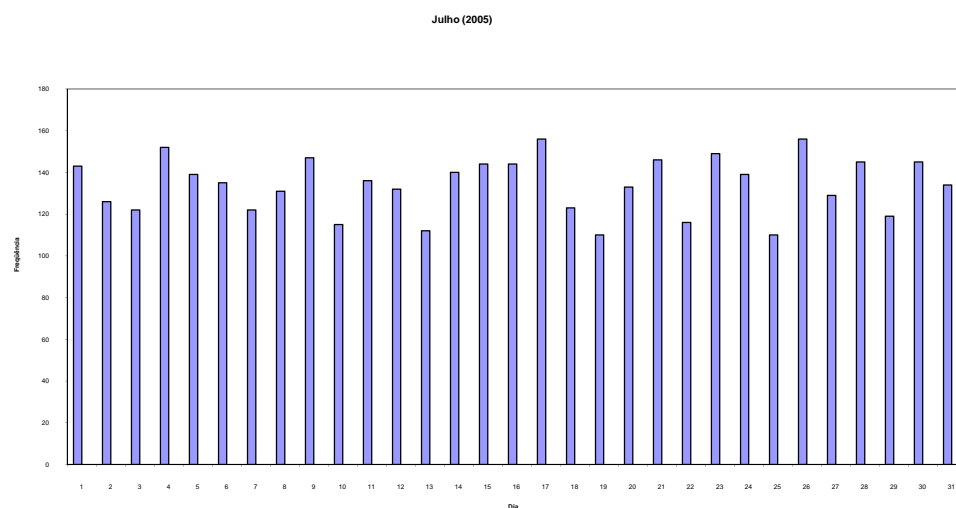


Figura C7 – Atendimentos de urgência e emergência realizados pelo SAMU_RP em julho de 2005.

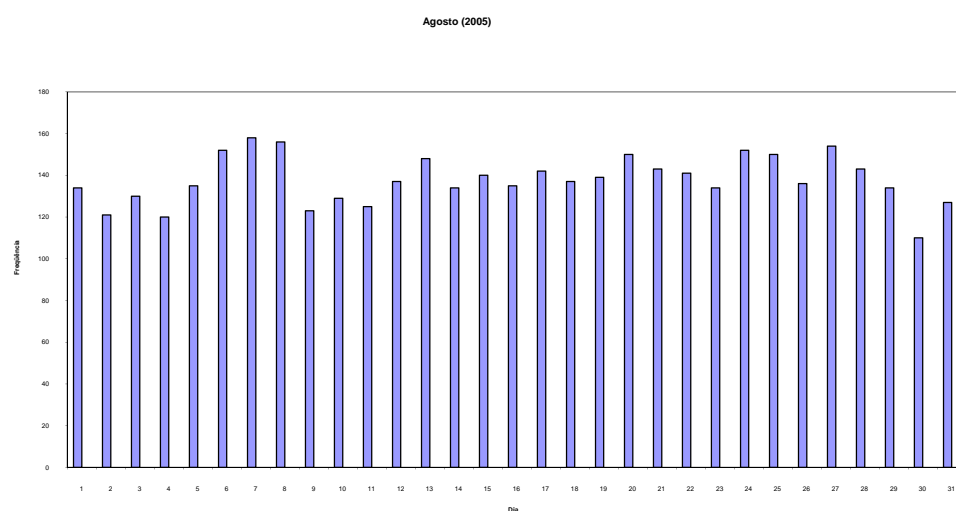


Figura C8 – Atendimentos de urgência e emergência realizados pelo SAMU_RP em agosto de 2005.

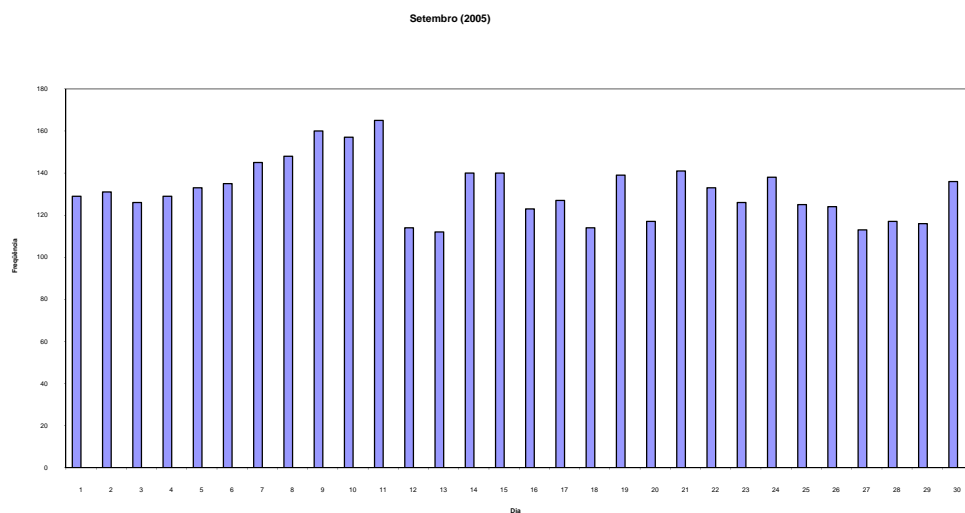


Figura C9 – Atendimentos de urgência e emergência realizados pelo SAMU_RP em setembro de 2005.

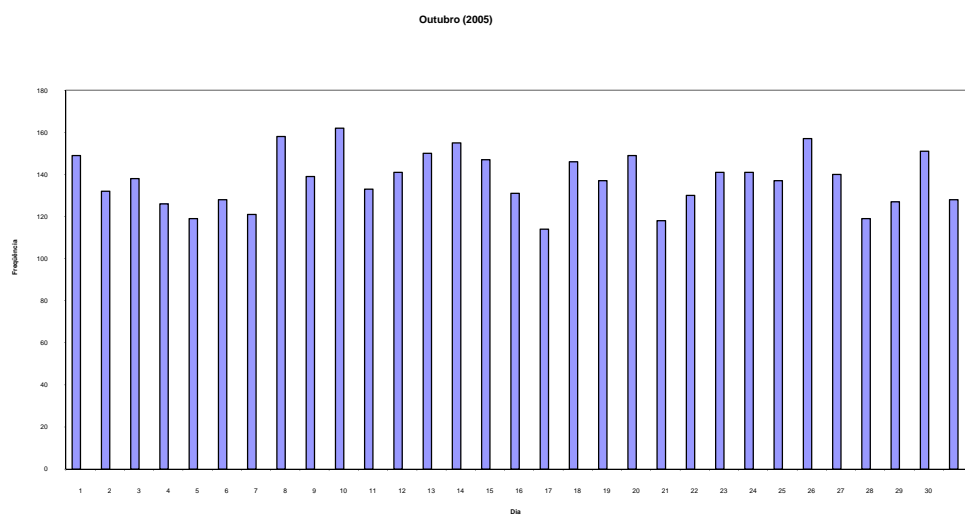


Figura C10 – Atendimentos de urgência e emergência realizados pelo SAMU_RP em outubro de 2005.

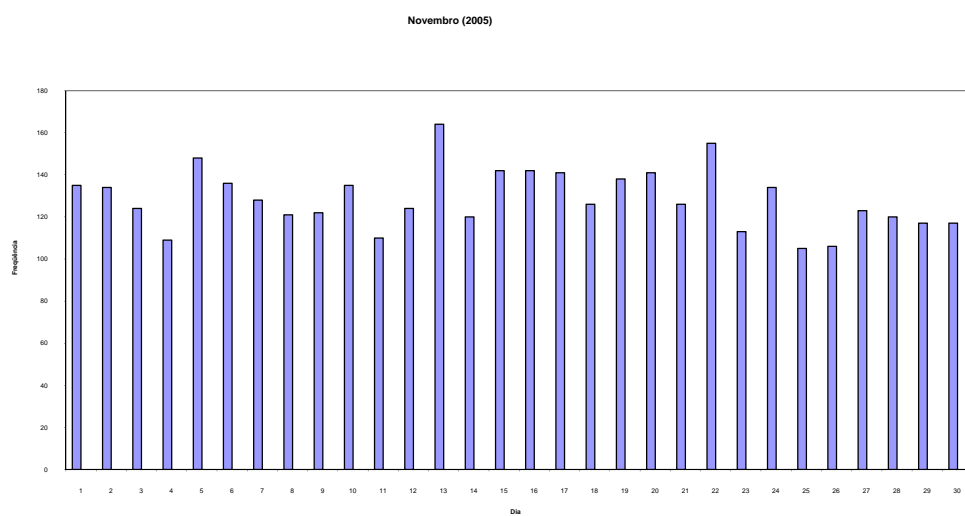


Figura C11 – Atendimentos de urgência e emergência realizados pelo SAMU_RP em novembro de 2005.

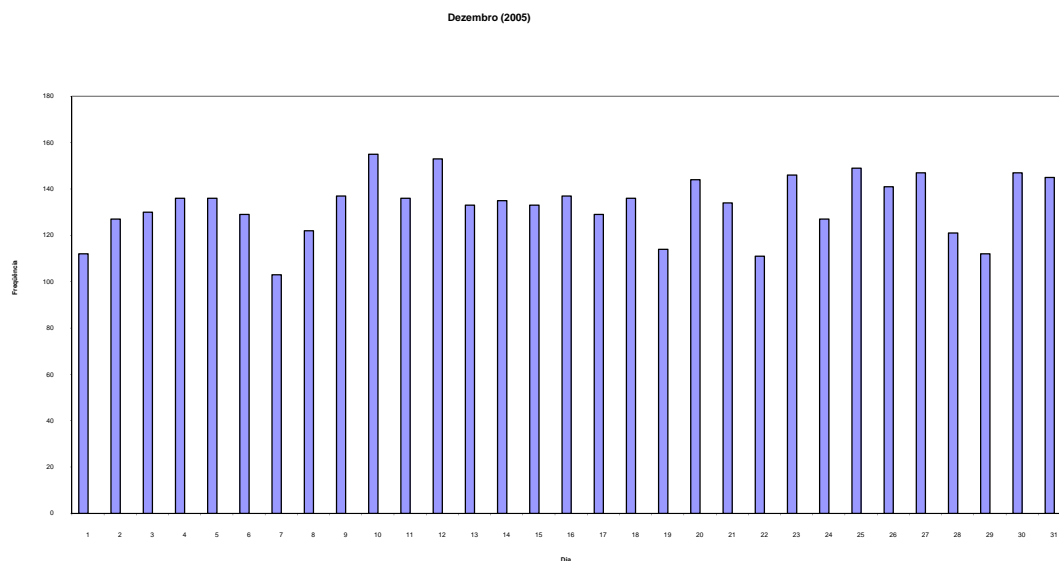


Figura C12 – Atendimentos de urgência e emergência realizados pelo SAMU_RP em dezembro de 2005.

As figuras C13 a C22, a seguir, mostram o número de chamados (frequência) ao longo das 24 horas dos dez dias escolhidos aleatoriamente do mês de agosto de 2005.

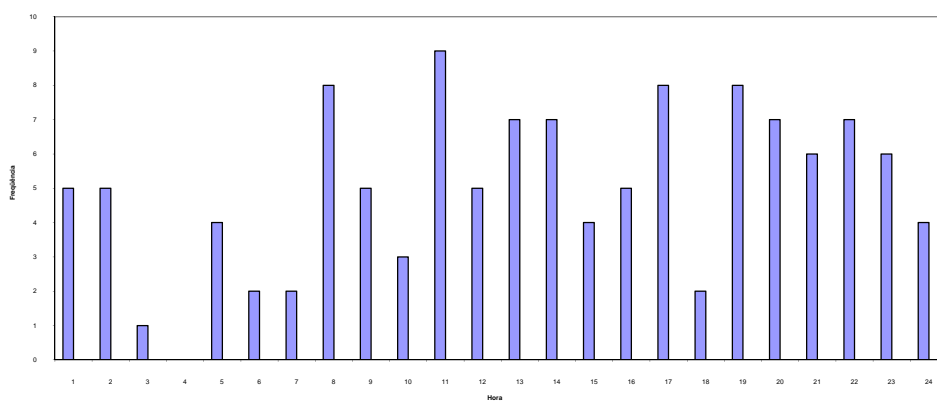


Figura C13 – Atendimentos realizados pelo SAMU_RP no 1º dia observado em agosto de 2005.

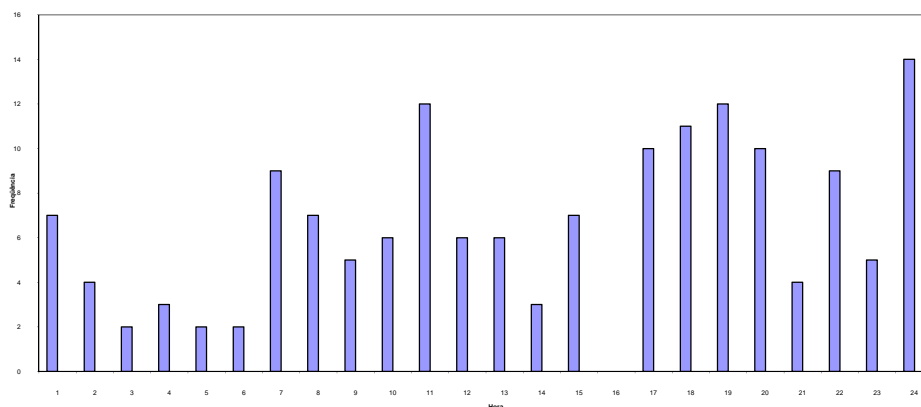


Figura C14 – Atendimentos realizados pelo SAMU_RP no 2º dia observado em agosto de 2005.

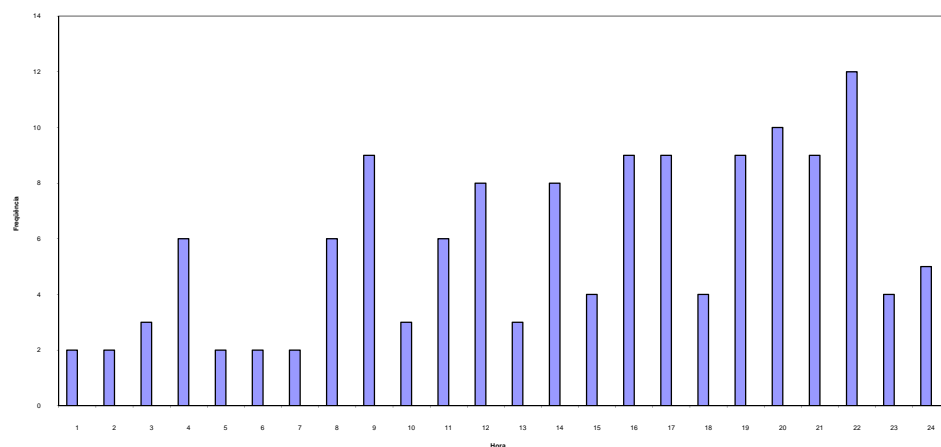


Figura C15 – Atendimentos realizados pelo SAMU_RP no 3º dia observado em agosto de 2005.

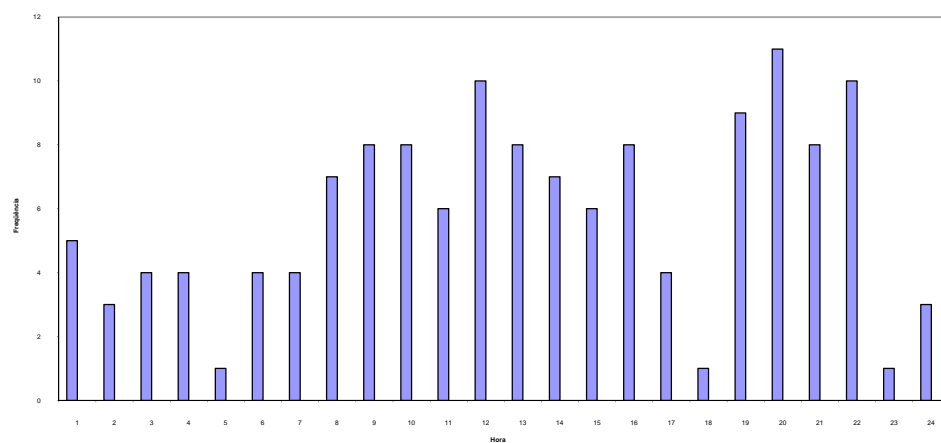


Figura C16 – Atendimentos realizados pelo SAMU_RP no 4º dia observado em agosto de 2005.

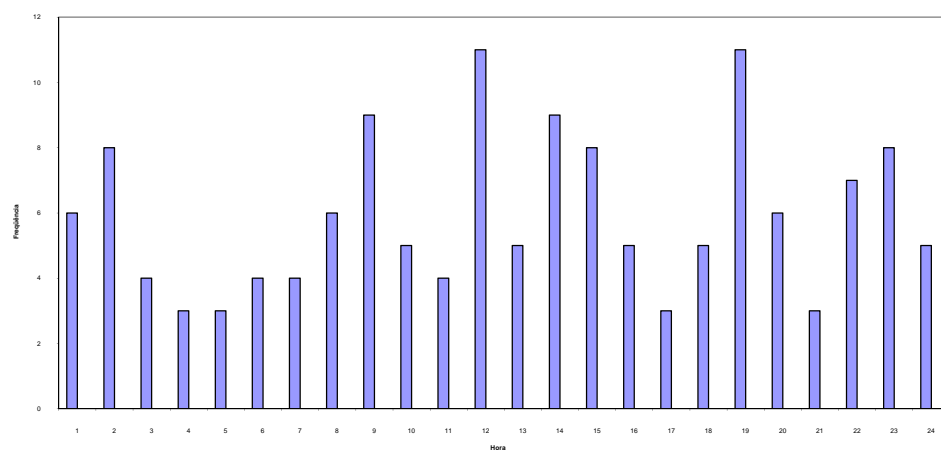


Figura C17 – Atendimentos realizados pelo SAMU_RP no 5º dia observado em agosto de 2005.

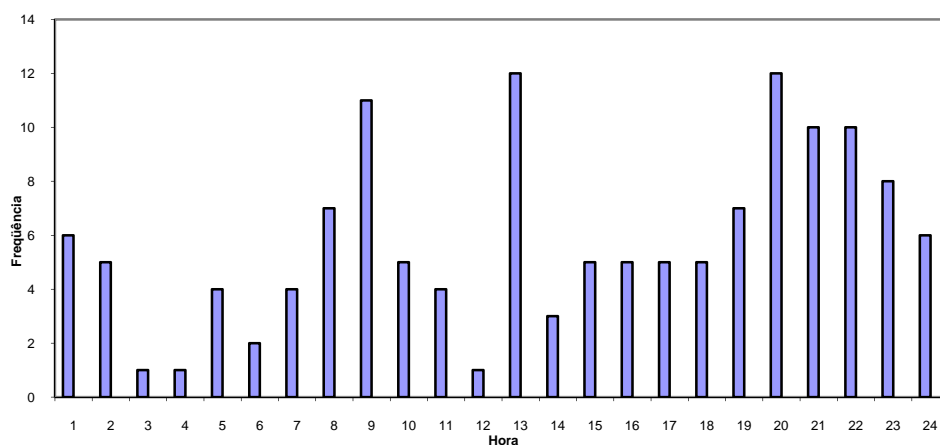


Figura C18 – Atendimentos realizados pelo SAMU_RP no 6º dia observado em agosto de 2005.

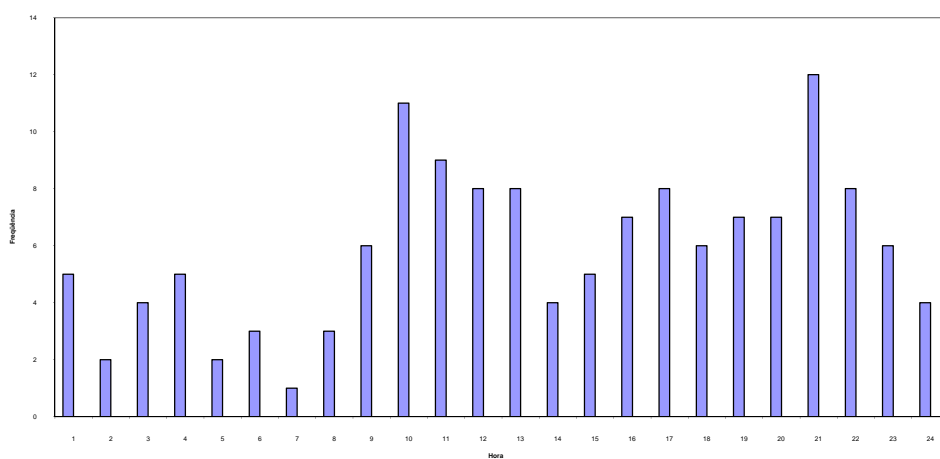


Figura C19 – Atendimentos realizados pelo SAMU_RP no 7º dia observado em agosto de 2005.

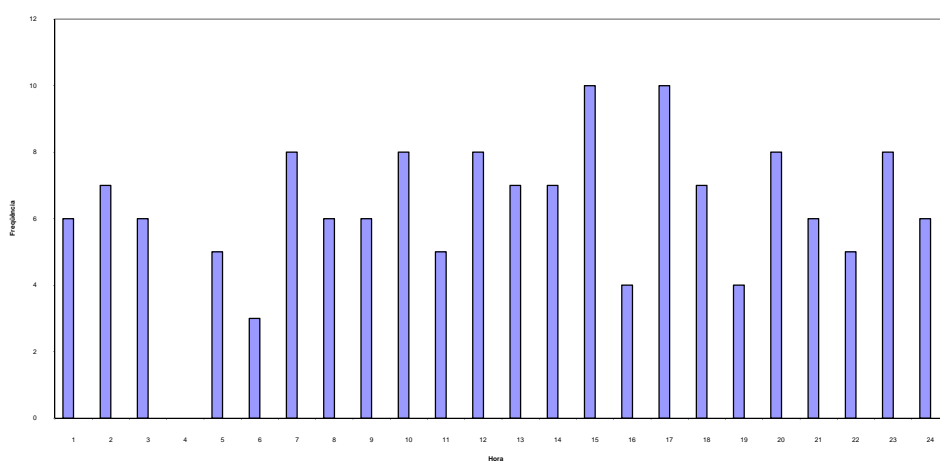


Figura C20 – Atendimentos realizados pelo SAMU_RP no 8º dia observado em agosto de 2005.

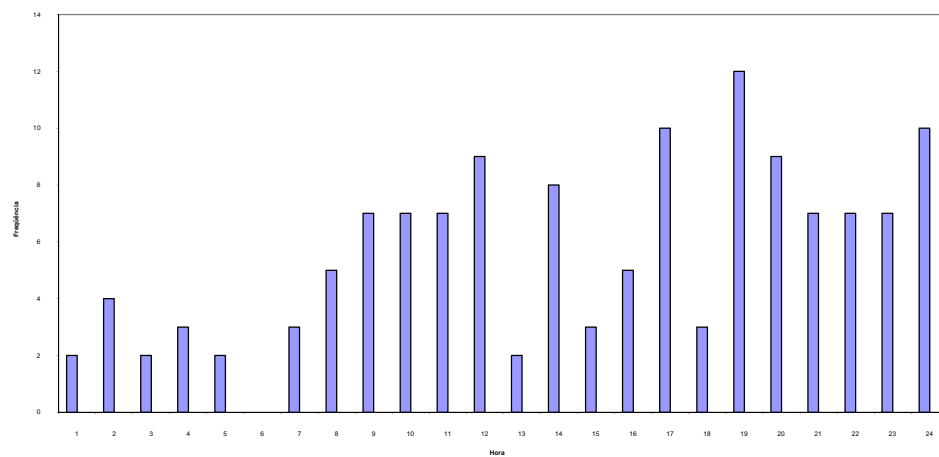


Figura C21 – Atendimentos realizados pelo SAMU_RP no 9º dia observado em agosto de 2005.

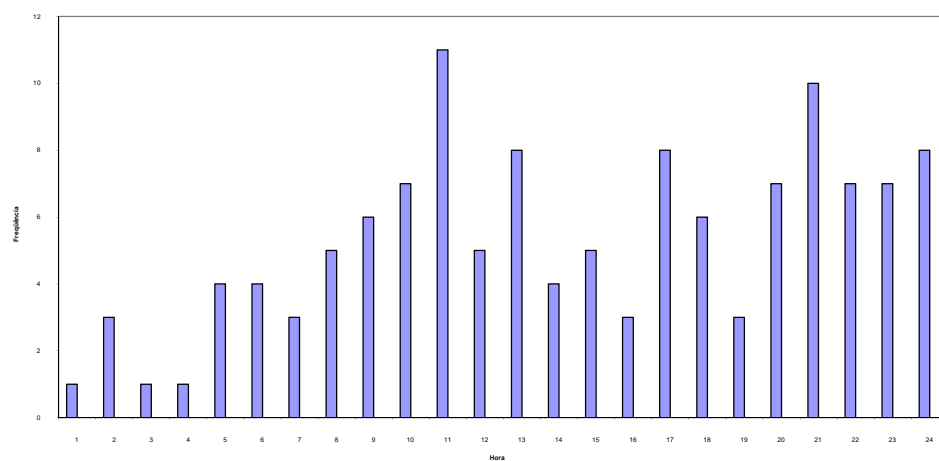


Figura C22 – Atendimentos realizados pelo SAMU_RP no 10º dia observado em agosto de 2005.

ANEXO D

O modelo de simulação do exemplo ilustrativo feito no software ARENA[®] representa um sistema com:

- ✓ 3 átomos contendo chamadas do tipo (prioridade): a, b e c.
- ✓ 3 servidores, sendo que o servidor 1 é um VSA (atende com prioridade chamadas do tipo a) e servidores 2 e 3 são servidores VSB's (atendem como primeiro e segundo servidores da lista de despacho chamadas do tipo b e c.
- ✓ fila com prioridade (chamados a, b e c) e com limite de tamanho de até 3 servidores.
- ✓ lista fixa de preferência de despacho para cada sub-átomo, sendo que servidor 1 pode atender chamadas do tipo a e b como terceiro servidor.

O modelo está dividido em três partes: processo de chegada, processo de atendimento e processo de fila e despacho de ambulâncias:

1. Processo de chegada:

O processo de chegada é simulado de forma separada para cada átomo 1, 2 e 3, e então são gerados os chamados dos 3 diferentes tipo em cada átomo. Neste modulo, definiu-se a distribuição do intervalo entre chamadas (utilizou-se a distribuição exponencial negativa) de cada sub-átomo: $1a$, $1b$, $1c$, $3c$.

Em seguida, também para cada sub-átomo, definiu-se os atributos de cada tipo de chamada em cada átomo: tipo de chamada (prioridade); lista de preferência de despacho: primeiro, segundo e terceiro servidor; tempo de viagem de cada servidor a este sub-átomo (segundo a ordem da lista de chegadas).

2. Processo de fila e despacho de servidores:

Após a chegada de um chamado no sistema, verifica-se a disponibilidade dos servidores da lista de despacho. Se todos os 3 servidores do sistema (seguindo a ordem da lista de preferência de despacho) estiverem ocupados, verifica-se o tamanho da fila, e se esta tem menos que 3 chamadas, a chamada que acaba de chegar espera em fila até que um servidor seja liberado. Caso a fila esteja cheia (com 3 chamadas) e os 3 servidores estejam ocupados, esta chamada é perdida para o sistema.

No caso de algum dos 3 servidores da lista de despacho estar livre, o servidor passa do estado livre para ocupado. Em seguida, simula-se a viagem deste servidor a chamada lendo o atributo relacionado aos tempos de viagem e a relação dos atributos (base (servidor) – átomo).

3. Processo de atendimento:

O processo de atendimento é simulado de forma separada para cada servidor 1, 2 e 3, definindo as distribuições dos tempos de serviço. Estas distribuições são exponenciais negativas, representando o tempo de atendimento no local e viagem de volta a base. O tempo de viagem da base ao átomo é descontado, pois já foi computado na etapa anterior. Após terminar o serviço (que compreende a viagem de volta a base), o servidor é liberado. A chamada (entidade) sai do sistema depois de terminado o serviço e liberado o servidor.

Ao longo da rodada do modelo de simulação, foram coletadas as estatísticas relacionadas a: tempo médio de viagem ao sistema, tempo de viagem de cada tipo de chamada, tempo de viagem de cada servidor, tempo de viagem de cada servidor a cada tipo de chamada, taxa de ocupação de cada servidor (carga de trabalho), tempo de fila para todas as chamadas e somente para chamadas que realmente esperam em fila, frequência de despacho de cada servidor a cada sub-átomo, e tempo médio de sistema.

O modelo de simulação do SAMU-RP feito no software ARENA[®] representa um sistema com:

- ✓ 5 átomos contendo chamadas do tipo (prioridade): a , b e c .
- ✓ 10 servidores, sendo que o servidor 1 é uma VSA (atende com prioridade as chamadas do tipo a) e os demais são servidores idênticos VSB (atendem como primeiro e segundo servidores da lista de despacho chamadas do tipo b e c).
- ✓ Fila com prioridade (a , b e c) e com limite de tamanho de até 5 servidores.
- ✓ Política de despacho:
 - Chamadas são atendidas com preferência pelos servidores localizados em seus átomos;
 - Se estes estiverem ocupados a chamada é atendida de forma aleatória por outro servidor *backup* (VSB) se houver algum disponível;
 - Caso somente a VSA estiver disponível a chamada é atendida por este;

O modelo está dividido em três partes: processo de chegada, processo de atendimento e processo de fila e despacho de ambulâncias. As entidades são as chamadas de emergência, e os recursos correspondem aos servidores (ambulâncias).

1. Processo de chegada:

O processo de chegada é simulado de forma separada para cada átomo, com o módulo CREATE são geradas as chamadas dos 3 diferentes tipos em cada átomo (gerando os sub-átomos). Neste módulo define-se a distribuição do intervalo entre chamadas (exponencial negativa em todos os sub-átomos) de cada sub-átomo: $1a$, $1b$, $1c$, ..., $5c$. Em seguida, também para cada sub-átomo utilizamos o módulo ASSIGN para definir os atributos de cada tipo de chamada em cada átomo, estes são:

- ✓ Tipo de chamada (prioridade);
- ✓ Servidores preferenciais;
- ✓ Instante de chegada.

2. Processo de fila e despacho de servidores:

Após a chegada de uma chamada no sistema (entidade), verifica-se primeiro se o tipo de chamada:

- ✓ Se a chamada é do tipo *a*, então verifica-se se o servidor VSA está disponível. Se estiver livre, este servidor passa a ser ocupado (é despachado). Se este servidor estiver ocupado, verifica-se a disponibilidade dos servidores preferenciais do átomo (origem desta chamada). Se houver dois servidores preferenciais no átomo e os dois estiverem disponíveis, escolhe-se de forma aleatória. Ou se, apenas um estiver ocupado, este é despachado. Se estes estiverem ocupados, verifica-se a disponibilidade dos servidores *backup*, e escolhe-se de forma aleatória entre os servidores *backup* disponíveis;
- ✓ Se a chamada é do tipo *b* ou *c*, verifica-se a disponibilidade dos servidores preferenciais do átomo (podem ser 1 ou 2 servidores dependendo do átomo). Se os dois servidores estiverem disponíveis, escolhe-se de forma aleatória entre os dois. Se somente um estiver disponível, este é despachado (se torna ocupado). Se os servidores preferenciais estiverem ocupados, escolhe-se de forma aleatória entre os servidores *backup* disponíveis. Se todos os servidores VSB estiverem ocupados, a chamada é atendida pela VSA caso esta estiver livre. A simulação destas regras de despacho é feita usando módulos DECIDE e SEIZE.
- ✓ Caso todos os servidores estiverem ocupados, a chamada entra em fila se esta tiver menos que 5 chamadas já em espera. Caso a fila esteja cheia, a chamada é perdida para o sistema (modulo DISPOSE).

Das distribuições que o Arena trabalha, a distribuição que passou no teste de todas as VSB's e melhor foram se adequou aos dados foi a Lognormal. Houve apenas cinco observações para o VSA durante o período observado, assim não foi possível obter uma distribuição a partir da amostra para esta ambulância. Em seguida simulamos a viagem deste servidor com a chamada utilizando os módulos DECIDE e LEAVE.

3. Processo de atendimento:

O processo de atendimento é simulado de forma separada para cada servidor, onde utilizamos o modulo DELAY para definir as distribuições de tempo de serviço. Este

intervalo de tempo inclui tempo de atendimento no local, tempo de viagem e outros tempos, durante o qual o servidor se encontra ocupado. Após terminar o serviço (que compreende a viagem de volta a base), o servidor é liberado (passa do estado ocupado a livre) através do modulo RELEASE. As distribuições dos tempos de serviço utilizadas foram:

- ✓ Ambu1: CONT(0.2, 0.6, 0.4, 0.63333, 0.6, 0.78333, 0.8, 0.9, 1.0, 1.2)
- ✓ Ambu 2 e 3 : LOGN(0.502255, 0.283752)
- ✓ Ambu 4 e 5: LOGN(0.75078, 0.434415)
- ✓ Ambu 6 e 7: LOGN(0.555286, 0.255442)
- ✓ Ambu 8: LOGN(0.640566, 0.339931)
- ✓ Ambu 9 e 10: LOGN(0.655178, 0.491563)

onde:

Cont = *Empirical Continuous Distribution* (Prob1, Value1, Prob2, Value2, . . . [,Stream]);
LOGN = *Lognormal*(μ, σ).

4. Calculo dos tempos de viagem:

No modelo hipercubo os tempos de viagem estão incluídos nos parâmetros da distribuição exponencial. Desta forma, no processo de validação deste modelo, não poderíamos apenas subtrair o tempo de viagem de cada entidade (ao ser servida pelo servidor) na etapa anterior. As estatísticas relacionadas ao tempo de viagem são computadas, após liberação dos servidores, utilizando-se os dados da matriz de tempos de viagem servidor-átomo e os módulos ASSIGN (para determinar o instante de partida), DELAY e RECORD. A chamada (entidade) sai do sistema através do modulo DISPOSE.

Ao longo da rodada do modelo de simulação, são coletadas estatísticas relacionadas a: tempo médio de viagem ao sistema, tempo de viagem de cada tipo de chamada, tempo de viagem de cada servidor, tempos médios de viagem de cada sub-átomo, taxa de ocupação de cada servidor (carga de trabalho), tempo de fila para todas as chamadas e somente para chamadas que realmente esperam em fila, tempos em fila para cada tipo de chamada, frequência de despacho de cada servidor a cada sub-átomo, e tempo médio de sistema. Para coletar estas estatísticas utilizamos o modulo RECORD.

Finalmente, utilizamos a função SETUP para definir o tamanho da simulação, o tempo de aquecimento (*warm up*). Por exemplo, esta simulação tem um *warm-up* de 3000 horas e o tempo total (com *warm-up*) de 103000, sendo que as estatísticas começam a ser coletadas após o tempo de *warm-up*.

ANEXO E

Tabelas dos resultados do modelo hipercubo para o cenário original

a) Período da manhã.

A probabilidade de encontrar o sistema vazio ($p[0000000000]$), a probabilidade de todos os servidores estarem ocupados ($P_{1111111111}$) é de 0,0001 e 0,0061, respectivamente. A probabilidade de fila no sistema (P_Q) é 0,0038 e o tempo de espera na fila (W_q) é de 0,1127.

Tempo médio de espera em fila (min.)	Sistema	a	b	c
Modelo	0,1	0,1	0,1	0,1

Tabela E1 – Tempo médio de espera na fila (manhã).

	Tempo médio de viagem no sistema	Tempo médio de resposta no sistema
Modelo	9,8	9,9

Tabela E2 – Tempo médio de viagem e resposta no sistema (manhã).

		Tempo médio de viagem (minutos)	Tempo médio de resposta (minutos)
Ambulância	Workload	Modelo	Modelo
1	0,16	9,8	9,9
2	0,42	8,9	9,0
3	0,42	8,9	9,0
4	0,53	11	11,1
5	0,53	11,3	11,4
6	0,42	9,3	9,4
7	0,43	9,1	9,2
8	0,46	9,9	10,0
9	0,39	9,6	9,7
10	0,39	10,2	10,3
VSA	0,16	9,8	9,9
VSB	0,44	9,8	9,9
Total	0,42	9,8	9,9

Tabela E3 – Tempo médio de viagem e resposta para cada ambulância (manhã).

	Tempo médio de viagem (minutos)	Tempo médio de resposta (minutos)
Sub-átomo	Modelo	Modelo
1a	9,1	9,2
1b	9	9,1
1c	8,8	8,9
2a	10,9	11,0
2b	10,5	10,6
2c	10,3	10,4
3a	9,1	9,2
3b	9	9,1
3c	9,8	9,9
4a	11,4	11,5
4b	12	12,1
4c	10,9	11,0
5a	9,4	9,5
5b	9,1	9,2
5c	9,5	9,6
Média a	10,0	10,0
Média b	9,9	10,0
Média c	9,9	10,0

Tabela E4 – Tempo médio de viagem e resposta em cada átomo (manhã).

	Tempo médio de viagem (minutos)	Tempo médio de resposta (minutos)
Átomos	Modelo	Modelo
a	9,7	9,8
b	9,6	9,7
c	9,9	10,0

Tabela E5 – Tempo médio de viagem e resposta para cada classe de usuário (manhã).

	Tempo médio de viagem			Tempo médio de resposta		
Ambulância	a	b	c	a	b	c
1	9,8	9,9	9,8	9,9	10,0	9,9
2	8,8	8,8	8,9	8,9	8,9	9,0
3	8,8	8,9	9	8,9	9,0	9,1
4	11,4	10,7	11,2	11,5	10,8	11,3
5	11,5	11,1	11,5	11,6	11,2	11,6
6	8,8	9,2	9,4	8,9	9,3	9,5
7	8,8	8,8	9,3	8,9	8,9	9,4
8	9,7	10,1	9,7	9,8	10,2	9,8
9	9,4	9,1	10,6	9,5	9,2	10,7
10	9,2	10,4	9,4	9,3	10,5	9,5

Tabela E6 – Tempo médio de viagem e resposta para ambulância/átomo (manhã).

b) Período da noite.

A probabilidade de encontrar o sistema vazio ($p[0000000000]$) e a probabilidade de todos os servidores estarem ocupados ($P_{1111111111}$) são de 0,0023 e

menor que 10^{-4} , respectivamente. A probabilidade de fila no sistema (P_Q) e o tempo de espera na fila (W_q) são menores que 10^{-4} .

Tempo médio de espera em fila (min.)	Sistema	a	b	c
Modelo	0,0	0,0	0,0	0,0

Tabela E7 – Tempo médio de espera na fila (noite).

	Tempo médio de viagem no sistema (min.)	Tempo médio de resposta no sistema (min.)
Modelo	9,8	9,8

Tabela E8 – Tempo médio de viagem e resposta no sistema (noite).

		Tempo Médio de Viagem (minutos)	Tempo Médio de Resposta (minutos)
Ambulância	Workload	Modelo	Modelo
1	0,02	9,7	9,7
2	0,09	9	9,0
3	0,1	9	9,0
4	0,3	11,6	11,6
5	0,3	11	11,0
6	0,15	8,8	8,8
7	0,15	8,9	8,9
8	0,19	10,3	10,3
9	0,23	9,7	9,7
VSA	0,02	9,7	9,7
VSB	0,19	9,8	9,8
Total	0,17	9,8	9,8

Tabela E9 – Tempo médio de viagem e resposta para cada ambulância (noite).

Noite	Tempo Médio de Viagem (minutos)	Tempo Médio de Resposta (minutos)
Átomos	Modelo	Modelo
<i>a</i>	9,6	9,6
<i>b</i>	9,7	9,7
<i>c</i>	10,1	10,1

Tabela E10 – Tempo médio de viagem e resposta para cada classe de usuário (noite).

	Tempo Médio de Viagem (minutos)	Tempo Médio de Resposta (minutos)
Sub-átomo	Modelo	Modelo
1a	8,7	8,7
1b	8,7	8,7
1c	8,7	8,7
2a	11,3	11,3
2b	11,3	11,3
2c	10,6	10,6
3a	8,7	8,7
3b	8,7	8,7
3c	9,5	9,5
4a	10,9	10,9
4b	11	11,0
4c	10,9	10,9
5a	9,7	9,7
5b	9	9,0
5c	9,6	9,6
Média a	9,9	9,9
Média b	9,7	9,7
Média c	9,9	9,9

Tabela E11 – Tempo médio de viagem e resposta em cada átomo (noite).

	Tempo médio de viagem - Modelo			Tempo médio de resposta - Modelo		
Ambulância	a	b	c	a	b	c
1	9,7	10	10	9,7	10,0	10,0
2	8,7	8,9	9,1	8,7	8,9	9,1
3	8,7	9,1	8,7	8,7	9,1	8,7
4	11,6	11,6	11,6	11,6	11,6	11,6
5	11,6	10,8	11,6	11,6	10,8	11,6
6	8,6	8,8	8,7	8,6	8,8	8,7
7	8,6	8,6	9,7	8,6	8,6	9,7
8	12,4	10,2	10,3	12,4	10,2	10,3
9	10	9,6	10	10,0	9,6	10,0

Tabela E12 – Tempo médio de viagem e resposta para ambulância/átomo (noite).

ANEXO F

Tabelas dos resultados do modelo hipercubo para o cenário 1.

a) Resultados para o período da manhã.

A probabilidade de todos os servidores estarem ocupados ($P_{iiiiiiii}$) é de 0,1048.

Manhã	Tempo médio de espera em fila (min.)	Sistema	a	b	c
	Cenário 1	11,3	3,6	4,2	13,4
	Simulação	11,2	3,5	4,2	13,4
Desvio	Minutos	0,038	0,078	0,029	0,044
	%	0,3	2,2	0,7	0,3

Tabela F1 – Tempo médio de espera na fila, cenário 1 (manhã).

Manhã		Tempo médio de viagem no sistema	Tempo médio de resposta no sistema
	Cenário 1	10,2	21,4
	Simulação	10,2	21,4
Desvio	Minutos	-0,032	0,01
	%	-0,3	0,0

Tabela F2 – Tempo médio de viagem e resposta no sistema, cenário 1 (manhã).

Manhã	Workload		Desvio	
Ambulância	Cenário 1	Simulação	Diferença	%
1	0,87	0,88	-0,003	-0,3
2	0,94	0,94	0,000	0,0
3	0,94	0,93	0,001	0,2
4	0,95	0,95	0,001	0,1
5	0,95	0,95	0,001	0,1
6	0,94	0,94	-0,001	-0,2
7	0,94	0,94	-0,001	-0,2
8	0,96	0,96	-0,003	-0,3
9	0,94	0,93	0,003	0,3
10	0,94	0,93	0,003	0,3
VSA	0,87	0,88	-0,003	-0,3
VSB	0,94	0,94	0,000	0,0
Total	0,94	0,94	0,000	0,0

Tabela F3 – *Workload*, cenário 1 (manhã).

Ambulância	Tempo médio de viagem (minutos)				Tempo médio de resposta (minutos)			
	Cenário 1	Simulação	Desvio Minutos	Desvio (%)	Cenário 1	Simulação	Desvio Minutos	Desvio (%)
1	10,2	10,1	0,036	0,4	11,3	11,6	-0,283	-2,4
2	10,1	10,0	0,061	0,6	11,3	11,8	-0,510	-4,3
3	10,1	10,0	0,060	0,6	11,3	11,8	-0,504	-4,3
4	10,1	9,4	0,716	7,6	12,0	18,9	-6,867	-36,4
5	10,2	9,4	0,746	7,9	12,0	19,2	-7,141	-37,3
6	10,3	10,5	-0,173	-1,7	11,1	9,6	1,519	15,9
7	10,2	10,5	-0,273	-2,6	11,0	8,6	2,370	27,5
8	10,4	11,2	-0,846	-7,6	10,4	3,7	6,746	183,5
9	10,0	10,5	-0,467	-4,5	10,8	6,8	4,033	59,6
10	10,2	10,5	-0,222	-2,1	11,0	9,1	1,940	21,3
VSA	10,2	10,1	0,036	0,4	11,3	11,6	-0,283	-2,4
VSB	10,2	10,2	-0,044	-0,2	11,2	11,0	0,176	25,1
Total	10,2	10,2	-0,036	-0,1	11,2	11,1	0,130	22,3

Tabela F4 – Tempo médio de viagem e resposta a cada ambulância, cenário 1.

Sub-átomo	Tempo médio de viagem (minutos)		Desvio		Tempo médio de resposta (minutos)		Desvio	
	Cenário 1	Simulação	Minutos	%	Cenário 1	Simulação	Minutos	%
1a	9,8	9,5	0,2531	2,7	13,4	13,0	0,3	2,5
1b	9,9	9,6	0,2462	2,6	14,1	13,8	0,3	2,0
1c	9,8	9,6	0,206	2,1	23,3	23,0	0,2	1,1
2a	9,4	9,4	-0,0343	-0,4	13,0	12,9	0,0	0,3
2b	9,4	9,4	-0,0064	-0,1	13,6	13,6	0,0	0,2
2c	9,3	9,4	-0,1027	-1,1	22,8	22,8	-0,1	-0,3
3a	10,4	10,4	0,0905	0,9	14,0	13,9	0,2	1,2
3b	10,4	10,2	0,1913	1,9	14,6	14,4	0,2	1,5
3c	10,4	10,3	0,1332	1,3	23,8	23,6	0,2	0,7
4a	11,4	11,8	-0,3445	-2,9	15,0	15,3	-0,3	-1,7
4b	11,4	11,8	-0,4216	-3,6	15,6	16,0	-0,4	-2,5
4c	11,6	11,8	-0,1951	-1,7	25,1	25,2	-0,2	-0,6
5a	10,1	9,8	0,2268	2,3	13,7	13,4	0,3	2,3
5b	10,1	9,8	0,2789	2,9	14,3	14,0	0,3	2,2
5c	10,0	9,8	0,2669	2,7	23,5	23,2	0,3	1,3
Média a	10,2	10,2	0,0	0,5	13,8	13,7	0,1	0,9
Média b	10,2	10,2	0,1	0,7	14,4	14,4	0,1	0,7
Média c	10,2	10,2	0,1	0,7	23,7	23,6	0,1	0,5

Tabela F5 – Tempo médio de viagem e resposta para cada sub-átomo, cenário 1 (manhã).

Manhã	Tempo Médio de Viagem (minutos)			Tempo Médio de Resposta (minutos)		
Átomos	Cenário 1	Simulação	Desvio (%)	Cenário 1	Simulação	Desvio (%)
a	10,1	10,4	-2,8	13,7	13,9	-1,6
b	10,1	10,2	-0,6	14,3	14,4	-0,2
c	10,2	10,2	-0,1	23,6	23,6	0,1

Tabela F6 – Tempo médio de resposta para cada classe de usuário, cenário 1 (manhã).

Manhã	Tempo médio de viagem			Tempo médio de resposta		
Ambulância	a	b	c	a	b	c
1	9,8	9,9	9,8	11,8	12,2	16,3
2	8,8	8,8	8,9	10,8	11,1	15,4
3	8,8	8,9	9	10,8	11,2	15,5
4	11,4	10,7	11,2	13,4	13,0	17,7
5	11,5	11,1	11,5	13,5	13,4	18,0
6	8,8	9,2	9,4	10,8	11,5	15,9
7	8,8	8,8	9,3	10,8	11,1	15,8
8	9,7	10,1	9,7	11,7	12,4	16,2
9	9,4	9,1	10,6	11,4	11,4	17,1
10	9,2	10,4	9,4	11,2	12,7	15,9

Tabela F7 – Tempo médio de viagem e resposta para ambulância/átomo, cenário 1.

b) Resultados para o período da noite.

A probabilidade de todos os servidores estarem ocupados ($P_{iiiiiiii}$) é 0,0652 e a probabilidade de fila é de 0,1127. O tempo de espera na fila (cenário alternativo) resultou em 1,4 minutos.

Noite	Tempo médio de espera em fila (min.)	Sistema	A	b	c
	Cenário 1	0,8	0,4	0,4	0,8
	Simulação	0,8	0,4	0,4	0,9
Desvio	Minutos	0,0	0,0	0,0	0,0
	%	0,0	-4,6	0,7	-0,1

Tabela F8 – Tempo médio de espera na fila, cenário 1 (noite).

Noite		Tempo médio de viagem no sistema	Tempo médio de resposta no sistema
	Cenário 1	9,9	10,7
	Simulação	10,0	10,8
Desvio	Minutos	-0,1	-0,1
	%	-0,9	-0,8

Tabela F9 – Tempo médio de viagem e resposta no sistema, cenário 1 (noite).

Noite	Workload		Desvio	
Ambulância	μ 's cenário original	Simulação	Diferença	%
1	0,30	0,31	-0,006	-1,9
2	0,56	0,56	0,003	0,5
3	0,56	0,56	0,003	0,4
4	0,73	0,74	-0,007	-0,9
5	0,74	0,74	0,003	0,4
6	0,62	0,62	0,003	0,6
7	0,62	0,62	0,002	0,4
8	0,70	0,70	0,001	0,1
9	0,64	0,64	0,000	0,0
VSA	0,3039	0,31	-0,006	-1,9
VSB	0,65	0,65	0,001	0,2
Total	0,61	0,61	0,000	-0,1

Tabela F10 – Workload, cenário 1 (noite).

Ambulância	Tempo médio de viagem (minutos)				Tempo médio de resposta (minutos)			
	Cenário 1	Simulação	Desvio Minutos	Desvio (%)	Cenário 1	Simulação	Desvio Minutos	Desvio (%)
1	10,1	10,0	0,1	0,7	10,9	10,8	0,1	0,6
2	9,7	9,6	0,1	0,8	10,5	10,4	0,1	0,8
3	9,5	9,6	-0,1	-1,0	10,3	10,4	-0,1	-0,9
4	9,8	10,1	-0,3	-2,8	10,6	10,9	-0,3	-2,6
5	10,3	10,1	0,1	1,4	11,1	10,9	0,1	1,3
6	10,2	10,0	0,3	2,5	11,0	10,7	0,3	2,4
7	9,6	10,0	-0,4	-4,1	10,3	10,7	-0,4	-3,8
8	10,6	10,8	-0,2	-2,3	11,4	11,6	-0,2	-2,1
9	9,6	9,9	-0,3	-2,6	10,4	10,7	-0,3	-2,4
VSA	10,1	10,0	0,1	0,7	10,9	10,8	0,1	0,6
VSB	9,9	10,0	-0,1	-1,0	10,7	10,8	-0,1	-0,9
Total	9,9	10,0	-0,1	-0,8	10,7	10,8	-0,1	-0,8

Tabela F11 – Tempo médio de viagem e resposta para cada ambulância, cenário 1 (noite).

Noite	Tempo Médio de Viagem (minutos)			Tempo Médio de Resposta (minutos)		
Átomos	μ 's cenário original	Simulação	Desvio (%)	μ 's cenário original	Simulação	Desvio (%)
a	9,7	9,7	-0,1	10,1	10,1	-0,1
b	9,9	9,9	0,3	10,3	10,3	0,3
c	9,9	10,0	-1,1	10,7	10,9	-1,0

Tabela F12 – Tempo médio de viagem e resposta a cada classe de usuário, cenário 1 (noite).

Observe, na tabela F13, que os átomos 4a e 5a não possuem valores calculados pela simulação, isso se deve ao fato de ter ocorrido apenas 1 chamada a cada 1000000 horas.

Sub-átomo	Tempo médio de viagem (minutos)		Desvio		Tempo médio de resposta (minutos)		Desvio	
	Cenário 1	Simulação	Minutos	%	Cenário 1	Simulação	Minutos	%
1a	10,0	8,9	1,0	11,7	10,3	9,3	1,0	11,0
1b	9,9	9,2	0,7	7,2	10,3	9,6	0,7	7,0
1c	10,3	9,2	1,1	11,6	11,1	10,1	1,1	10,6
2a	10,1	9,6	0,5	4,8	10,4	10,0	0,4	4,5
2b	9,8	10,2	-0,4	-3,8	10,2	10,6	-0,4	-3,7
2c	9,6	10,1	-0,6	-5,5	10,4	11,0	-0,6	-5,1
3a	11,1	10,6	0,5	5,0	11,5	11,0	0,5	4,6
3b	11,5	9,5	2,0	20,5	11,9	9,9	2,0	19,7
3c	10,7	9,6	1,2	12,2	11,6	10,4	1,2	11,2
4a	9,8	---	---	---	10,2	---	---	---
4b	10,0	11,2	-1,2	-11,0	10,4	11,6	-1,2	-10,6
4c	9,7	11,2	-1,5	-13,7	10,5	12,1	-1,5	-12,8
5a	9,0	---	---	---	9,3	---	---	---
5b	9,2	9,9	-0,7	-7,1	9,6	10,3	-0,7	-6,8
5c	9,2	9,9	-0,7	-6,7	10,1	10,7	-0,7	-6,1
Média a	10,0	5,8	0,4	4,3	10,4	6,1	0,4	4,0
Média b	10,1	10,0	0,1	1,2	10,5	10,4	0,1	1,1
Média c	9,9	10,0	-0,1	-0,4	10,7	10,8	-0,1	-0,4

Tabela F13 – Tempo médio de viagem e resposta para cada sub-átomo, cenário 1 (noite).

Ambulância	Tempo médio de viagem - Noite			Tempo médio de resposta		
	a	b	c	a	b	c
1	9,7	10	10	10,3	10,7	11,5
2	8,7	8,9	9,1	9,3	9,6	10,6
3	8,7	9,1	8,7	9,3	9,8	10,2
4	11,6	11,6	11,6	12,2	12,3	13,1
5	11,6	10,8	11,6	12,2	11,5	13,1
6	8,6	8,8	8,7	9,2	9,5	10,2
7	8,6	8,6	9,7	9,2	9,3	11,2
8	12,4	10,2	10,3	13,0	10,9	11,8
9	10	9,6	10	10,6	10,3	11,5

Tabela F14 – Tempo médio de viagem e resposta para cada ambulância/átomo, cenário 1 (noite).

ANEXO G

Equações de balanço do exemplo ilustrativo referente ao hipercubo do Capítulo 4.

1) Quando todos os servidores estão livres.

Somente o estado $B = \{000\}$ corresponde ao estado na qual todos os servidores estão livres. Seus vértices adjacentes são $\{001\}, \{010\}, \{100\}$, veja a tabela 4.1 e a figura 4.3. Seja P_B a probabilidade de o sistema estar no estado B , a equação de equilíbrio do estado $B = \{000\}$, é dada por:

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

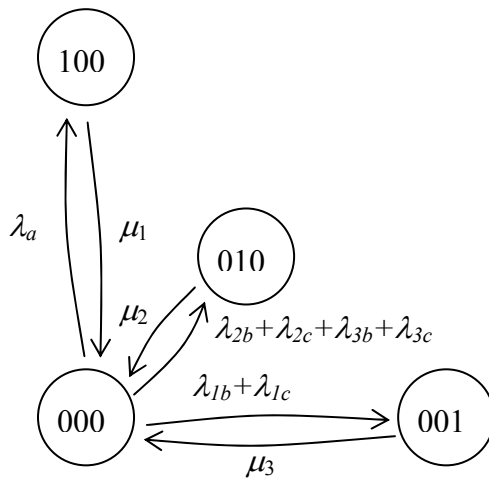


Figura G1 – Vértice $\{000\}$ e seus adjacentes.

Na equação 1 o lado esquerdo da igualdade corresponde ao fluxo para fora do estado $B = \{000\}$. Veja a Tabela 4.1 e a Figura 4.3 que as transições de estado para seus vértices adjacentes são:

➤ $\{000\} \rightarrow \{100\}$, quando ocorre uma chamada nos subátomos $1a$, $2a$ ou $3a$,

com taxa λ_a ;

➤ $\{000\} \rightarrow \{010\}$, quando ocorre uma chamada nos subátomos $2b$, $2c$, $3b$ ou

$3c$, com taxa $\lambda_{2b} + \lambda_{2c} + \lambda_{3b} + \lambda_{3c}$;

➤ $\{000\} \rightarrow \{001\}$, quando ocorre uma chamada nos subátomos $1b$ ou $1c$,

com taxa $\lambda_{1b} + \lambda_{1c}$.

Nota-se que $\lambda = \lambda_a + (\lambda_{2b} + \lambda_{2c} + \lambda_{3b} + \lambda_{3c}) + (\lambda_{1b} + \lambda_{1c})$. O lado direito da igualdade (1) corresponde ao fluxo para dentro do estado $B = \{000\}$. As transições para o vértice $\{000\}$, saindo de seus adjacentes (veja a Tabela 4.1 e a Figura 4.3), são:

➤ $\{100\} \rightarrow \{000\}$, quando ocorre um término de serviço do servidor 1, com

taxa ;

➤ $\{010\} \rightarrow \{000\}$, quando ocorre um término de serviço do servidor 2, com

taxa μ_2 ;

➤ $\{001\} \rightarrow \{000\}$, quando ocorre um término de serviço do servidor 3, com

taxa μ_3 .

2) Quando há um servidor ocupado no sistema.

Quando há um servidor ocupado no sistema têm-se três possíveis estados para o sistema com um servidor ocupado $\{100\}, \{010\}, \{001\}$. Eles serão analisados a seguir.

I. Para o estado $\{100\}$, com seus vértices adjacentes $\{000\}, \{110\}, \{101\}$, veja a tabela 4.1 e a figura 4.4. A equação de equilíbrio do estado $\{100\}$ é dada por:

$$(\lambda + \mu_1)P_{\{100\}} = \mu_2P_{\{110\}} + \mu_3P_{\{101\}} + \lambda_aP_{\{000\}} \quad (2)$$

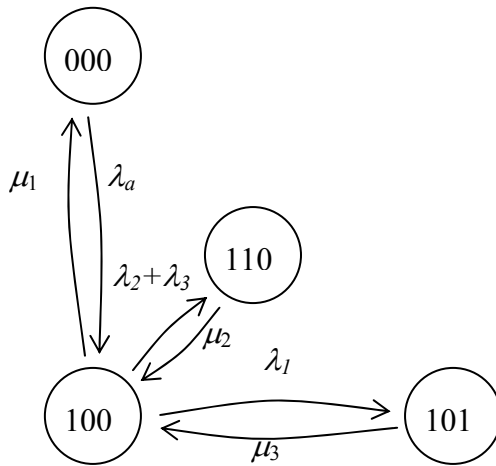


Figura G2 – Vértice $\{100\}$ e seus adjacentes.

O lado esquerdo da equação (2) corresponde ao fluxo para fora do estado $B = \{100\}$. Na Tabela 4.1 e na Figura 4.4, as transições de estado para os vértices adjacentes a $\{100\}$ são:

- $\{100\} \rightarrow \{000\}$, quando ocorre um término de serviço do servidor 1, com taxa μ_1 ;
- $\{100\} \rightarrow \{101\}$, quando ocorre uma chamada nos subátomos $1a$, $1b$ ou $1c$, com taxa λ_1 .
- $\{100\} \rightarrow \{110\}$, quando ocorre uma chamada nos subátomos $2a$, $2b$, $2c$, $3a$, $3b$ ou $3c$, com taxa $(\lambda_2 + \lambda_3)$.

O lado direito da equação (2) corresponde ao fluxo para dentro do estado $B = \{100\}$. Na Tabela 4.1 e na Figura 4.4, as transições de estado para os vértices adjacentes a $\{100\}$ são:

- $\{000\} \rightarrow \{100\}$, quando ocorre uma chamada nos subátomos $1a$, $2a$ ou $3a$, com taxa λ_a ;
- $\{101\} \rightarrow \{100\}$, quando ocorre um término de serviço do servidor 3, com taxa μ_3 ;

➤ $\{110\} \rightarrow \{100\}$, quando ocorre um término de serviço do servidor 2, com taxa .

II. O estado $\{010\}$, com seus vértices adjacentes $\{000\}, \{110\}, \{011\}$, veja a Tabela 4.1 e a Figura 4.5. A equação de equilíbrio do estado $\{010\}$ é dada por:

$$(\lambda + \mu_2)P_{\{010\}} = \mu_1 P_{\{110\}} + \mu_3 P_{\{011\}} + (\lambda_{2b} + \lambda_{2c} + \lambda_{3b} + \lambda_{3c})P_{\{000\}} \quad (3)$$

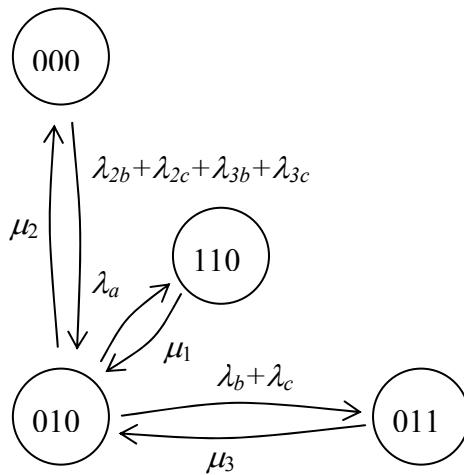


Figura G3 – Vértice $\{010\}$ e seus adjacentes.

O lado esquerdo da equação (3) corresponde ao fluxo para fora do estado $B = \{010\}$. Na Tabela 4.1 e na Figura 4.5, as transições de estado para os vértices adjacentes a $\{010\}$ são:

➤ $\{010\} \rightarrow \{000\}$, quando ocorre um término de serviço do servidor 2, com taxa μ_2 ;

➤ $\{010\} \rightarrow \{110\}$, quando ocorre uma chamada nos subátomos $1a$, $2a$ ou $3a$, com taxa λ_a ;

➤ $\{010\} \rightarrow \{011\}$, quando ocorre uma chamada nos subátomos $1b$, $1c$, $2b$, $2c$, $3b$ ou $3c$ (preferencial ou *backup*), com taxa $(\lambda_b + \lambda_c)$.

O lado direito da equação (3) corresponde ao fluxo para dentro do estado

$B = \{010\}$. Na Tabela 4.1 e na Figura 4.5 as transições de estado para os vértices adjacentes a $\{010\}$ são:

- $\{000\} \rightarrow \{010\}$, quando ocorre uma chamada nos subátomos $2b, 2c, 3b$ ou $3c$, com taxa $(\lambda_{2b} + \lambda_{2c} + \lambda_{3b} + \lambda_{3c})$;
- $\{110\} \rightarrow \{010\}$, quando ocorre um término de serviço do servidor 1, com taxa μ_1 ;
- $\{011\} \rightarrow \{010\}$, quando ocorre um término de serviço do servidor 3, com taxa μ_3 .

III. O estado $\{001\}$, com seus vértices adjacentes $\{000\}, \{101\}, \{011\}$, veja a Tabela 4.1 e a Figura 4.6. A equação de equilíbrio do estado $\{001\}$ é dada por:

$$(\lambda + \mu_3)P_{\{001\}} = (\lambda_{1b} + \lambda_{1c})P_{\{000\}} + \mu_1 P_{\{101\}} + \mu_2 P_{\{011\}} \quad (4)$$

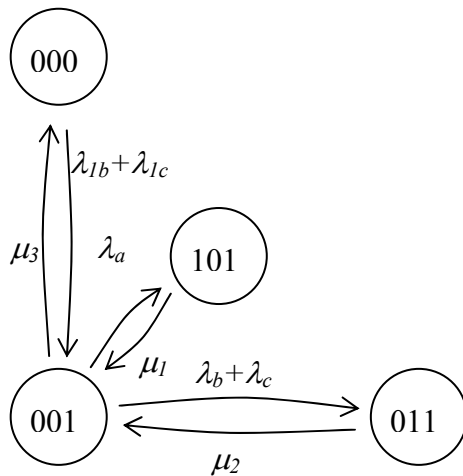


Figura G4 – Vértice $\{001\}$ e seus adjacentes.

O lado esquerdo da equação (4) corresponde ao fluxo para fora do estado $B = \{001\}$. Na Tabela 4.1 e na Figura 4.6, as transições de estado para os vértices

adjacentes a $\{001\}$ são:

- $\{001\} \rightarrow \{000\}$, quando ocorre um término de serviço do servidor 3, com taxa μ_3 ;
- $\{001\} \rightarrow \{011\}$, quando ocorre uma chamada nos subátomos $1b$, $1c$, $2b$, $2c$, $3b$ ou $3c$ (preferencial ou *backup*), com taxa $(\lambda_b + \lambda_c)$;
- $\{001\} \rightarrow \{101\}$, quando ocorre uma chamada nos subátomos $1a$, $2a$ ou $3a$, com taxa λ_a .

O lado direito da equação (4) corresponde ao fluxo para dentro do estado $B = \{001\}$. Na Tabela 4.1 e na Figura 4.6 as transições de estado para os vértices adjacentes a $\{001\}$ são:

- $\{000\} \rightarrow \{001\}$, quando ocorre uma chamada nos subátomos $1b$ ou $1c$, com taxa $(\lambda_{1b} + \lambda_{1c})$;
- $\{011\} \rightarrow \{001\}$, quando ocorre um término de serviço do servidor 2, com taxa μ_2 ;
- $\{101\} \rightarrow \{001\}$, quando ocorre um término de serviço do servidor 1, com taxa μ_1 .

3) Quando há dois servidores ocupados no sistema.

Quando há dois servidores ocupados no sistema, têm-se três possíveis estados, $\{110\}$, $\{101\}$, $\{011\}$, que serão analisados a seguir.

I. O estado $\{110\}$, com seus vértices adjacentes $\{111\}$, $\{010\}$, $\{100\}$, é mostrado na Figura 4.7. A equação de equilíbrio do estado $\{110\}$ é dada por:

$$(\lambda + \mu_1 + \mu_2)P_{\{110\}} = \mu_3P_{\{111\}} + \lambda_aP_{\{010\}} + (\lambda_2 + \lambda_3)P_{\{100\}} \quad (5)$$

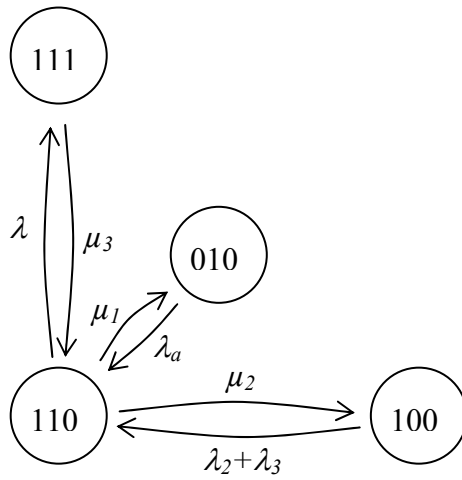


Figura G5 – Vértice $\{110\}$ e seus adjacentes.

O lado esquerdo da equação (5) corresponde ao fluxo para fora do estado $B = \{110\}$. Na Tabela 4.1 e na Figura 4.7, as transições de estado para os vértices adjacentes a $\{110\}$ são:

- $\{110\} \rightarrow \{111\}$, quando ocorre uma chamada em qualquer átomo de qualquer prioridade, com taxa λ ;
- $\{110\} \rightarrow \{010\}$, quando ocorre um término de serviço do servidor 1, com taxa μ_1 ;
- $\{110\} \rightarrow \{100\}$, quando ocorre um término de serviço do servidor 2, com taxa μ_2 .

O lado direito da equação (5) corresponde ao fluxo para dentro do estado $B = \{110\}$. Na Tabela 4.1 e na Figura 4.7 as transições de estado para os vértices adjacentes a $\{110\}$ são:

- $\{111\} \rightarrow \{110\}$, quando ocorre um término de serviço do servidor 3, com taxa μ_3 ;
- $\{010\} \rightarrow \{110\}$, quando ocorre uma chamada nos subátomos $1a$, $2a$ ou $3a$, com taxa λ_a ;

$\{100\} \rightarrow \{110\}$, quando ocorre uma chamada nos subátomos $2a$, $2b$, $2c$, $3a$, $3b$ ou $3c$, com taxa $(\lambda_2 + \lambda_3)$.

1) Quando todos os servidores estão ocupados.

Somente o estado $B = \{111\}$ corresponde ao estado onde todos os servidores estão ocupados. Seus vértices adjacentes são $\{011\}, \{010\}, \{110\}$ e $S_4 = \{a, b, c\}$, veja a Figura G6. Este último estado S_4 é uma fila que surge quando temos todos os servidores ocupados. Os sub-estados $\{a\}$, $\{b\}$ e $\{c\}$ de S_4 , representam os estados onde há 1 usuário em fila com prioridade a , b e c , respectivamente. A equação de equilíbrio do estado $B = \{111\}$ é dada por:

$$(\lambda + \mu)P_{\{111\}} = \lambda P_{\{011\}} + \lambda P_{\{101\}} + \lambda P_{\{110\}} + \mu(P_{\{a\}} + P_{\{b\}} + P_{\{c\}}) \quad (6)$$

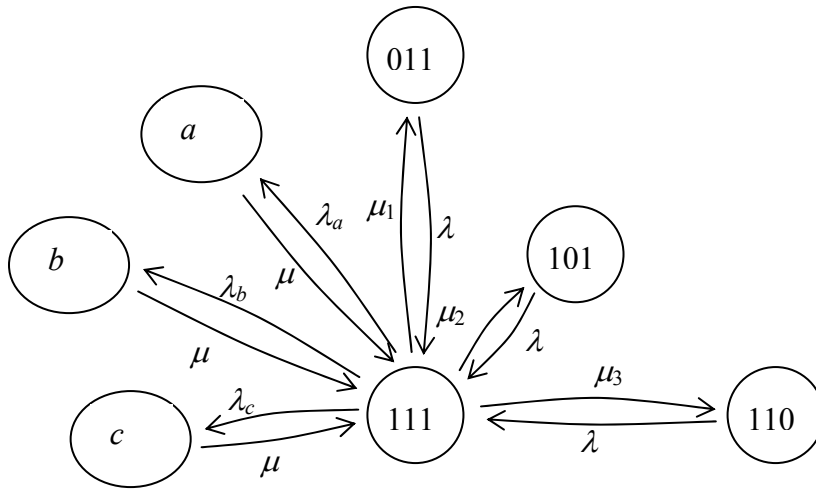


Figura G6 – Vértice $\{101\}$ e seus adjacentes.

O lado esquerdo da equação (6) corresponde ao fluxo para fora do estado $B = \{111\}$. Na Tabela 4.1 e na Figura G6, as transições de estado para os vértices adjacentes a $\{111\}$ são:

➤ $\{111\} \rightarrow \{011\}$, quando ocorre um término de serviço do servidor 1, com taxa μ_1 ;

- $\{111\} \rightarrow \{101\}$, quando ocorre um término de serviço do servidor 2, com taxa μ_2 ;
- $\{111\} \rightarrow \{110\}$, quando ocorre um término de serviço do servidor 3, com taxa μ_3 ;
- $\{111\} \rightarrow \{a,b,c\}$, quando ocorre uma chamada em qualquer subátomo de qualquer prioridade, com taxa λ .

O lado direito da equação (6) corresponde ao fluxo para dentro do estado $B = \{111\}$. Na Tabela 4.1 e na Figura G6, as transições de estado para os vértices adjacentes a $\{111\}$ são:

- $\{011\} \rightarrow \{111\}$, quando ocorre uma chamada nos subátomos $1a$, $2a$ ou $3a$, com taxa λ_a ;
- $\{101\} \rightarrow \{111\}$, quando ocorre uma chamada em qualquer subátomo de qualquer prioridade, com taxa λ ;
- $\{110\} \rightarrow \{111\}$, quando ocorre uma chamada em qualquer subátomo de qualquer prioridade, com taxa λ ;
- $\{a,b,c\} \rightarrow \{111\}$, quando ocorre um término do serviço de qualquer servidor, com taxa μ .

ANEXO H

Equações de balanço referente à fila do exemplo ilustrativo do Capítulo 4.

2) Quando há um usuário em fila.

I. O estado $\{a\}$, com seus vértices adjacentes $\{1,1,1\}, \{aa\}, \{ab\}$ e $\{ac\}$, é mostrado na Figura H1. A equação de equilíbrio do estado $\{a\}$ é dada por:

$$(\lambda + \mu)P_{\{a\}} = \mu P_{\{aa\}} + \lambda_a P_{\{111\}} \quad (7)$$

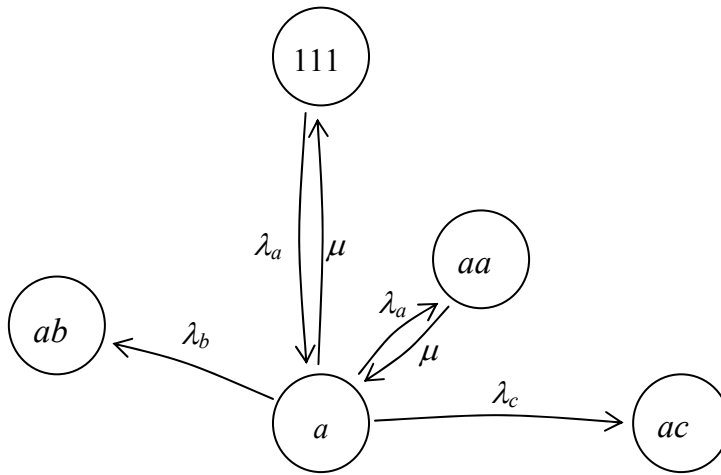


Figura H1 – Vértice $\{a\}$ e seus adjacentes.

O lado esquerdo da equação (7) corresponde ao fluxo para fora do estado $B = \{a\}$. Veja a equação (7) e a Figura H1, nas quais as transições de estado para os vértices adjacentes a $\{a\}$ são:

- $\{a\} \rightarrow \{aa\}$, quando ocorre uma chamada nos subátomos $1a$, $2a$ ou $3a$, com taxa λ_a ;
- $\{a\} \rightarrow \{ab\}$, quando ocorre uma chamada nos subátomos $1b$, $2b$ ou $3b$, com taxa λ_b ;
- $\{a\} \rightarrow \{ac\}$, quando ocorre uma chamada nos subátomos $1c$, $2c$, ou $3c$, com taxa λ_c ;

➤ $\{a\} \rightarrow \{111\}$, quando ocorre um término de serviço, com taxa μ .

O lado direito da equação (7) corresponde ao fluxo para dentro do estado $B = \{a\}$. Lembre-se que a chamada a tem prioridade de atendimento sobre os chamados b e c , por isso não há transição de $\{ab\} \rightarrow \{b\}$ e de $\{ac\} \rightarrow \{a\}$. Veja a equação (7) e a Figura H1, nas quais as transições de estado para os vértices adjacentes a $\{a\}$ são:

➤ $\{aa\} \rightarrow \{a\}$, quando ocorre um término de serviço, com taxa μ ;

➤ $\{111\} \rightarrow \{a\}$, quando ocorre uma chamada nos subátomos $1a$, $2a$ ou $3a$,

com taxa λ_a .

II. O estado $\{b\}$, com seus vértices adjacentes $\{111\}, \{bb\}, \{bc\}, \{ab\}$, é mostrado na Figura H2. A equação de equilíbrio do estado $\{b\}$ é dada por:

$$(\lambda + \mu)P_{\{b\}} = \lambda_b P_{\{111\}} + \mu P_{\{ab\}} + \mu P_{\{bb\}} \quad (8)$$

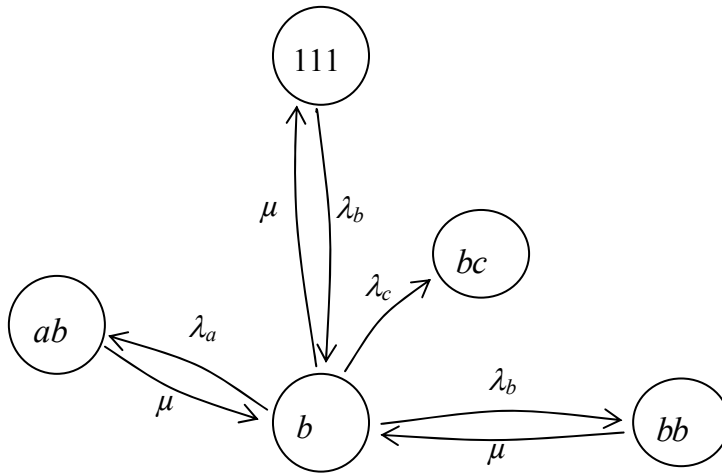


Figura H2 – Vértice $\{b\}$ e seus adjacentes.

O lado esquerdo da equação (8) corresponde ao fluxo para fora do estado $B = \{b\}$. Pode-se notar que a chamada b tem prioridade de atendimento sobre a chamada c , por isso não há transição de $\{bc\} \rightarrow \{b\}$. Na equação (8) e na Figura H2, as transições de estado para os vértices adjacentes a $\{b\}$ são:

- $\{b\} \rightarrow \{bb\}$, quando ocorre uma chamada nos subátomos $1b$, $2b$ ou $3b$, com taxa λ_b ;
- $\{b\} \rightarrow \{bc\}$, quando ocorre uma chamada nos subátomos $1c$, $2c$, ou $3c$, com taxa λ_c ;
- $\{b\} \rightarrow \{ab\}$, quando ocorre uma chamada nos subátomos $1a$, $2a$ ou $3a$, com taxa λ_a ;
- $\{b\} \rightarrow \{111\}$, quando ocorre um término de serviço, com taxa μ .

O lado direito da equação (8) corresponde ao fluxo para dentro do estado $B = \{b\}$. Veja a equação (8) e a Figura H2, nas quais as transições de estado para os vértices adjacentes a $\{b\}$ são:

- $\{ab\} \rightarrow \{b\}$, quando ocorre um término de serviço, com taxa μ ;
- $\{bb\} \rightarrow \{b\}$, quando ocorre um término de serviço, com taxa μ ;
- $\{111\} \rightarrow \{b\}$, quando ocorre uma chamada nos subátomos $1b$, $2b$ ou $3b$, com taxa λ_b ;

III. O estado $\{c\}$, com seus vértices adjacentes $\{111\}, \{cc\}, \{bc\}, \{ac\}$, é mostrado na Figura H3. A equação de equilíbrio do estado $\{c\}$ é dada por:

$$(\lambda + \mu)P_{\{c\}} = \lambda_c P_{\{111\}} + \mu P_{\{cc\}} + \mu P_{\{ac\}} + \mu P_{\{bc\}} \quad (9)$$

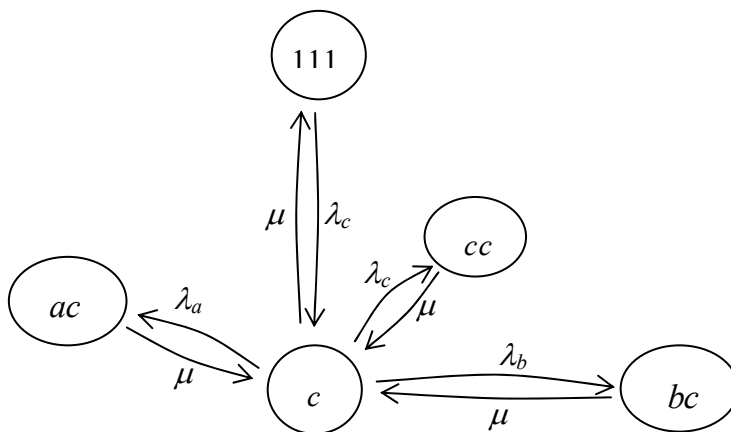


Figura H3 – Vértice $\{c\}$ e seus adjacentes.

O lado esquerdo da equação (9) corresponde ao fluxo para fora do estado $B = \{c\}$. Na equação (9) e na Figura H3, as transições de estado para os vértices adjacentes a $\{c\}$ são:

- $\{c\} \rightarrow \{cc\}$, quando ocorre uma chamada nos subátomos $1b$, $2b$ ou $3b$, com taxa λ_c ;
- $\{c\} \rightarrow \{bc\}$, quando ocorre uma chamada nos subátomos $1c$, $2c$, ou $3c$, com taxa λ_b ;
- $\{c\} \rightarrow \{ac\}$, quando ocorre uma chamada nos subátomos $1a$, $2a$ ou $3a$, com taxa λ_a ;
- $\{c\} \rightarrow \{111\}$, quando ocorre um término de serviço, com taxa μ .

O lado direito da equação (9) corresponde ao fluxo para dentro do estado $B = \{c\}$. Na equação (9) e na Figura H3, as transições de estado para os vértices adjacentes a $\{c\}$ são:

- $\{ac\} \rightarrow \{c\}$, quando ocorre um término de serviço, com taxa μ ;
- $\{cc\} \rightarrow \{c\}$, quando ocorre um término de serviço, com taxa μ ;
- $\{bc\} \rightarrow \{c\}$, quando ocorre um término de serviço, com taxa μ ;
- $\{111\} \rightarrow \{c\}$, quando ocorre uma chamada nos subátomos $1c$, $2c$, ou $3c$, com taxa λ_c .

3) Quando há dois usuários em fila:

I. O estado $\{aa\}$, com seus vértices adjacentes $\{a\}$, $\{aac\}$, $\{aab\}$, $\{aaa\}$, é mostrado na Figura H4. A equação de equilíbrio do estado $\{aa\}$ é dada por:

$$(\lambda + \mu)P_{\{aa\}} = \lambda_a P_{\{a\}} + \mu P_{\{aaa\}} \quad (10)$$

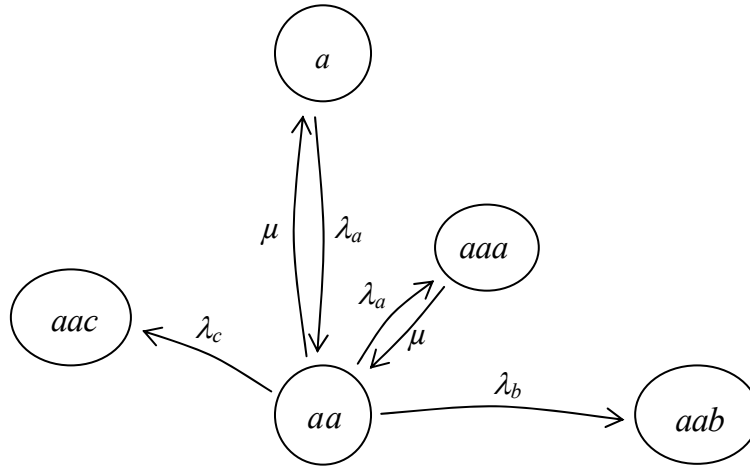


Figura H4– Vértice $\{aa\}$ e seus adjacentes.

O lado esquerdo da equação (10) corresponde ao fluxo para fora do estado $B = \{aa\}$. Na equação (10) e na Figura H4, as transições de estado para os vértices adjacentes a $\{aa\}$ são:

- $\{aa\} \rightarrow \{a\}$, quando ocorre um término de serviço, com taxa μ ;
- $\{aa\} \rightarrow \{aac\}$, quando ocorre uma chamada nos subátomos $1c$, $2c$, ou $3c$, com taxa λ_c ;
- $\{aa\} \rightarrow \{aab\}$, quando ocorre uma chamada nos subátomos $1b$, $2b$ ou $3b$, com taxa λ_b ;
- $\{aa\} \rightarrow \{aaa\}$, quando ocorre uma chamada nos subátomos $1a$, $2a$ ou $3a$, com taxa λ_a .

O lado direito da equação (10) corresponde ao fluxo para dentro do estado $B = \{aa\}$. Na equação (10) e na Figura H4, as transições de estado para os vértices adjacentes a $\{aa\}$ são:

- $\{a\} \rightarrow \{aa\}$, quando ocorre uma chamada nos subátomos $1a$, $2a$ ou $3a$, com taxa λ_a ;
- $\{aaa\} \rightarrow \{aa\}$, quando ocorre um término de serviço, com taxa μ .

II. O estado $\{bc\}$, com seus vértices adjacentes $\{b\}$, $\{c\}$, $\{bbc\}$, $\{bcc\}$, $\{abc\}$,

é mostrado na Figura H5. A equação de equilíbrio do estado $\{bc\}$ é dada por:

$$(\lambda + \mu)P_{\{bc\}} = \lambda_c P_{\{b\}} + \lambda_b P_{\{c\}} + \mu P_{\{abc\}} + \mu P_{\{bbc\}} \quad (11)$$

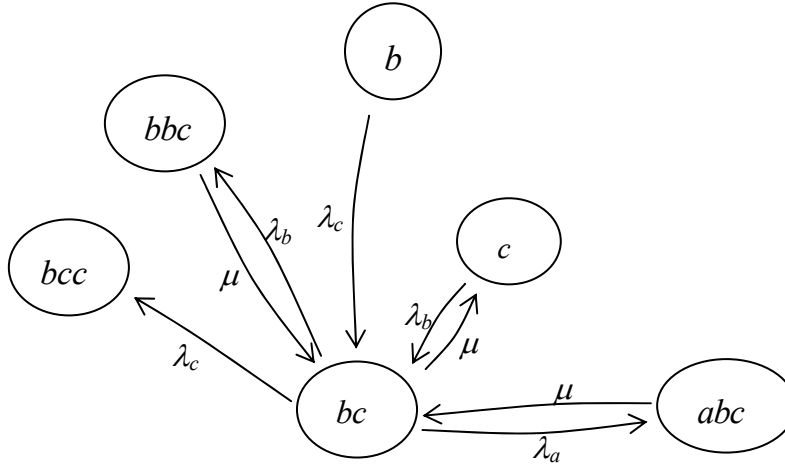


Figura H5 – Vértice $\{bc\}$ e seus adjacentes.

O lado esquerdo da equação (11) corresponde ao fluxo para fora do estado $B = \{bc\}$. Na equação (11) e na Figura H5, as transições de estado para os vértices adjacentes a $\{bc\}$ são:

- $\{bc\} \rightarrow \{c\}$, quando ocorre um término de serviço, com taxa μ ;
- $\{bc\} \rightarrow \{bbc\}$, quando ocorre uma chamada nos subátomos $1b$, $2b$ ou $3b$, com taxa λ_b ;
- $\{bc\} \rightarrow \{bcc\}$, quando ocorre uma chamada nos subátomos $1c$, $2c$, ou $3c$, com taxa λ_c ;
- $\{bc\} \rightarrow \{abc\}$, quando ocorre uma chamada nos subátomos $1a$, $2a$ ou $3a$, com taxa λ_a .

O lado direito da equação (11) corresponde ao fluxo para dentro do estado $B = \{bc\}$. Na equação (11) e na Figura H5, as transições de estado para os vértices adjacentes a $\{bc\}$ são:

- $\{b\} \rightarrow \{bc\}$, quando ocorre uma chamada nos subátomos $1c$, $2c$, ou $3c$, com taxa λ_c ;
- $\{c\} \rightarrow \{bc\}$, quando ocorre uma chamada nos subátomos $1b$, $2b$ ou $3b$, com taxa λ_b ;
- $\{bbc\} \rightarrow \{bc\}$, quando ocorre um término de serviço, com taxa μ .
- $\{abc\} \rightarrow \{bc\}$, quando ocorre um término de serviço, com taxa μ .

III. O estado $\{bb\}$, com seus vértices adjacentes $\{b\}, \{abb\}, \{bbb\}, \{bbc\}$, é mostrado na Figura H6. A equação de equilíbrio do estado $\{bb\}$ é dada por:

$$(\lambda + \mu)P_{\{bb\}} = \lambda_b P_{\{b\}} + \mu P_{\{abb\}} + \mu P_{\{bbb\}} \quad (12)$$

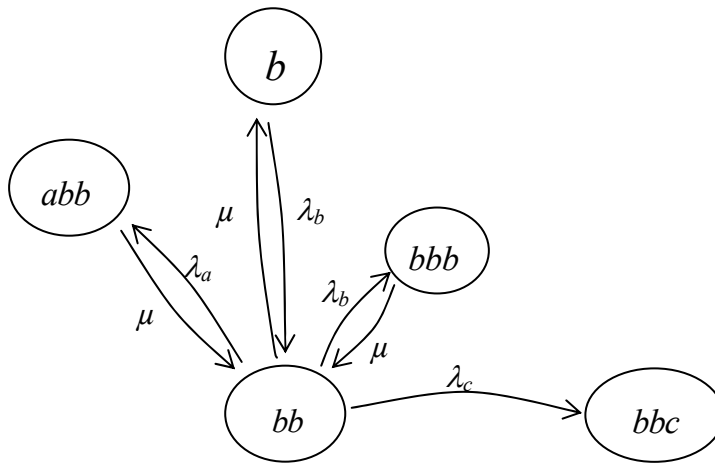


Figura H6 – Vértice $\{bb\}$ e seus adjacentes.

O lado esquerdo da equação (12) corresponde ao fluxo para fora do estado $B = \{bb\}$. Na equação (12) e na Figura H6, as transições de estado para os vértices adjacentes a $\{bb\}$ são:

- $\{bb\} \rightarrow \{b\}$, quando ocorre um término de serviço, com taxa μ ;
- $\{bb\} \rightarrow \{bbb\}$, quando ocorre uma chamada nos subátomos $1b$, $2b$ ou $3b$,

com taxa λ_b ;

- $\{bb\} \rightarrow \{bbc\}$, quando ocorre uma chamada nos subátomos $1c$, $2c$, ou $3c$,

com taxa λ_c ;

- $\{bb\} \rightarrow \{abb\}$, quando ocorre uma chamada nos subátomos $1a$, $2a$ ou $3a$,

com taxa λ_a .

O lado direito da equação (12) corresponde ao fluxo para dentro do estado $B=\{bb\}$. Na equação (12) e na Figura H6, as transições de estado para os vértices adjacentes a $\{bb\}$ são:

- $\{b\} \rightarrow \{bb\}$, quando ocorre uma chamada nos subátomos $1b$, $2b$ ou $3b$,

com taxa λ_b ;

- $\{bbb\} \rightarrow \{bb\}$, quando ocorre um término de serviço, com taxa μ ;

- $\{abb\} \rightarrow \{bb\}$, quando ocorre um término de serviço, com taxa μ .

IV. O estado $\{cc\}$, com seus vértices adjacentes $\{c\}, \{acc\}, \{bcc\}, \{ccc\}$, é mostrado na Figura H7. A equação de equilíbrio do estado $\{cc\}$ é dada por:

$$(\lambda + \mu)P_{\{cc\}} = \lambda_c P_{\{c\}} + \mu P_{\{bcc\}} + \mu P_{\{acc\}} + \mu P_{\{ccc\}} \quad (13)$$

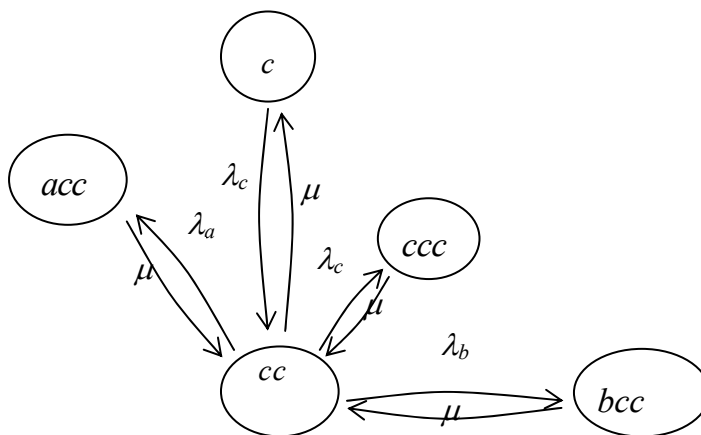


Figura H7 – Vértice $\{cc\}$ e seus adjacentes.

O lado esquerdo da equação (13) corresponde ao fluxo para fora do estado $B = \{cc\}$. Na equação (13) e na Figura H7, as transições de estado para os vértices adjacentes a $\{cc\}$ são:

- $\{cc\} \rightarrow \{c\}$, quando ocorre um término de serviço, com taxa μ ;
- $\{cc\} \rightarrow \{ccc\}$, quando ocorre uma chamada nos subátomos $1c$, $2c$, ou $3c$, com taxa λ_c ;
- $\{cc\} \rightarrow \{bcc\}$, quando ocorre uma chamada nos subátomos $1b$, $2b$ ou $3b$, com taxa λ_b ;
- $\{cc\} \rightarrow \{acc\}$, quando ocorre uma chamada nos subátomos $1a$, $2a$ ou $3a$, com taxa λ_a .

O lado direito da equação (13) corresponde ao fluxo para dentro do estado $B = \{cc\}$. Na equação (13) e na Figura H7, as transições de estado para os vértices adjacentes a $\{cc\}$ são:

- $\{c\} \rightarrow \{cc\}$, quando ocorre uma chamada nos subátomos $1c$, $2c$, ou $3c$, com taxa λ_c ;
- $\{ccc\} \rightarrow \{cc\}$, quando ocorre um término de serviço, com taxa μ ;
- $\{bcc\} \rightarrow \{cc\}$, quando ocorre um término de serviço, com taxa μ ;
- $\{acc\} \rightarrow \{cc\}$, quando ocorre um término de serviço, com taxa μ .

4) Quando há três usuários em fila.

O modelo não permite mais que três usuários em fila, o usuário que chega no sistema quando há três usuários é considerado uma perda.

I. O estado $\{aaa\}$, com seu vértice adjacente $\{aa\}$, é mostrado na Figura H8. A equação de equilíbrio do estado $\{aaa\}$ é dada por:

$$\mu P_{\{aaa\}} = \lambda_a P_{\{aa\}} \quad (14)$$

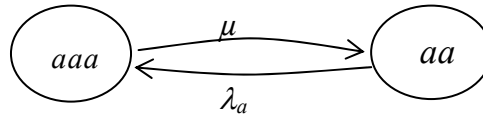


Figura H8 – Vértice $\{aaa\}$ e seus adjacentes.

O lado esquerdo da equação (14) corresponde ao fluxo para fora do estado $B = \{aaa\}$. Na equação (14) e na Figura H8, a transição de estado para o vértice adjacente a $\{aaa\}$ é:

➤ $\{aaa\} \rightarrow \{aa\}$, quando ocorre um término de serviço, com taxa μ .

O lado direito da equação (14) corresponde ao fluxo para dentro do estado $B = \{aaa\}$. Na equação (14) e na Figura H8, a transição de estado para o vértice adjacente a $\{aaa\}$ é:

➤ $\{aa\} \rightarrow \{aaa\}$, quando ocorre uma chamada nos subátomos $1a$, $2a$ ou $3a$, com taxa λ_a ;

II. O estado $\{aab\}$, com seus vértices adjacentes $\{aa\}, \{ab\}$, é mostrado na Figura H9. A equação de equilíbrio do estado $\{aab\}$ é dada por:

$$\mu P_{\{aab\}} = \lambda_a P_{\{ab\}} + \lambda_b P_{\{aa\}} \quad (15)$$

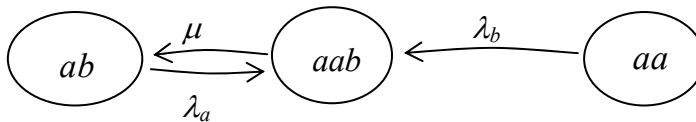


Figura H9 – Vértice $\{aab\}$ e seus adjacentes.

O lado esquerdo da equação (15) corresponde ao fluxo para fora do estado

$B = \{aab\}$. Na equação (15) e na Figura H9, a transição de estado para o vértice adjacente a $\{aab\}$ é:

➤ $\{aab\} \rightarrow \{ab\}$, quando ocorre um término de serviço, com taxa μ .

O lado direito da equação (15) corresponde ao fluxo para dentro do estado $B = \{aab\}$. Na equação (15) e na Figura H9, a transição de estado para os vértices adjacentes a $\{aab\}$ são:

➤ $\{ab\} \rightarrow \{aab\}$, quando ocorre uma chamada nos subátomos $1a$, $2a$ ou $3a$, com taxa λ_a ;

➤ $\{aa\} \rightarrow \{aab\}$, quando ocorre uma chamada nos subátomos $1b$, $2b$ ou $3b$, com taxa λ_b ;

III. O estado $\{aac\}$, com seus vértices adjacentes $\{aa\}, \{ac\}$, é mostrado na Figura H9. A equação de equilíbrio do estado $\{aac\}$ é dada por:

$$\mu P_{\{aac\}} = \lambda_a P_{\{ac\}} + \lambda_c P_{\{aa\}} \quad (16)$$

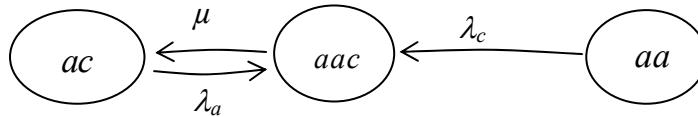


Figura H9 – Vértice $\{aac\}$ e seus adjacentes.

O lado esquerdo da equação (16) corresponde ao fluxo para fora do estado $B = \{aac\}$. Na equação (16) e na Figura H9, a transição de estado para o vértice adjacente a $\{aac\}$ é:

➤ $\{aac\} \rightarrow \{ac\}$, quando ocorre um término de serviço, com taxa μ .

O lado direito da equação (16) corresponde ao fluxo para dentro do estado $B = \{aac\}$. Na equação (16) e na Figura H9, a transição de estado para os vértices

adjacentes a $\{aac\}$ são:

➤ $\{ac\} \rightarrow \{aac\}$, quando ocorre uma chamada nos subátomos $1a$, $2a$ ou $3a$, com taxa λ_a ;

➤ $\{aa\} \rightarrow \{aac\}$, quando ocorre uma chamada nos subátomos $1c$, $2c$, ou $3c$, com taxa λ_c ;

IV. O estado $\{abb\}$, com seus vértices adjacentes $\{ab\}, \{bb\}$, é mostrado na Figura H10. A equação de equilíbrio do estado $\{abb\}$ é dada por:

$$\mu P_{\{abb\}} = \lambda_b P_{\{ab\}} + \lambda_a P_{\{bb\}} \quad (17)$$

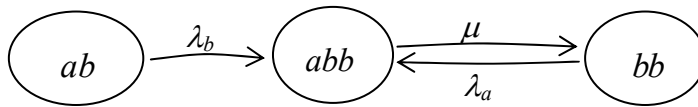


Figura H10 – Vértice $\{abb\}$ e seus adjacentes.

O lado esquerdo da equação (17) corresponde ao fluxo para fora do estado $B = \{abb\}$. Na equação (17) e na Figura H10, a transição de estado para o vértice adjacente a $\{abb\}$ é:

➤ $\{abb\} \rightarrow \{bb\}$, quando ocorre um término de serviço, com taxa μ .

O lado direito da equação (17) corresponde ao fluxo para dentro do estado $B = \{abb\}$. Na equação (17) e na Figura H10, a transição de estado para os vértices adjacentes a $\{abb\}$ são:

➤ $\{ab\} \rightarrow \{abb\}$, quando ocorre uma chamada nos subátomos $1b$, $2b$ ou $3b$, com taxa λ_b ;

➤ $\{bb\} \rightarrow \{abb\}$, quando ocorre uma chamada nos subátomos $1a$, $2a$ ou $3a$, com taxa λ_a ;

V. O estado $\{abc\}$, com seus vértices adjacentes $\{ab\}, \{ac\}, \{bc\}$, é mostrado na Figura H11. A equação de equilíbrio do estado $\{abc\}$ é dada por:

$$\mu P_{\{abc\}} = \lambda_a P_{\{bc\}} + \lambda_b P_{\{ac\}} + \lambda_c P_{\{ab\}} \quad (18)$$

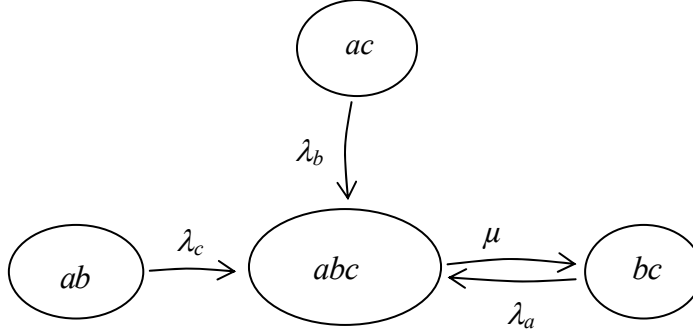


Figura H11 – Vértice $\{abc\}$ e seus adjacentes.

O lado esquerdo da equação (18) corresponde ao fluxo para fora do estado $B = \{abc\}$. Na equação (18) e na Figura H11, a transição de estado para o vértice adjacente a $\{abc\}$ é:

- $\{abc\} \rightarrow \{bc\}$, quando ocorre um término de serviço, com taxa μ .

O lado direito da equação (18) corresponde ao fluxo para dentro do estado $B = \{abc\}$. Na equação (18) e na Figura H11, a transição de estado para os vértices adjacentes a $\{abc\}$ são:

- $\{ab\} \rightarrow \{abc\}$, quando ocorre uma chamada nos subátomos $1c$, $2c$, ou $3c$, com taxa λ_c ;
- $\{ac\} \rightarrow \{abc\}$, quando ocorre uma chamada nos subátomos $1b$, $2b$ ou $3b$, com taxa λ_b ;
- $\{bc\} \rightarrow \{abc\}$, quando ocorre uma chamada nos subátomos $1a$, $2a$ ou $3a$, com taxa λ_a ;

VI. O estado $\{acc\}$, com seus vértices adjacentes $\{ac\}, \{cc\}$, é mostrado na Figura H12. A equação de equilíbrio do estado $\{acc\}$ é dada por:

$$\mu P_{\{acc\}} = \lambda_a P_{\{cc\}} + \lambda_c P_{\{ac\}} \quad (19)$$

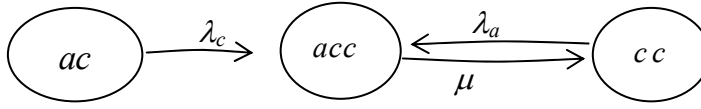


Figura H12– Vértice $\{acc\}$ e seus adjacentes.

O lado esquerdo da equação (19) corresponde ao fluxo para fora do estado $B = \{acc\}$. Na equação (19) e na Figura H12, a transição de estado para o vértice adjacente a $\{acc\}$ é:

- $\{acc\} \rightarrow \{cc\}$, quando ocorre um término de serviço, com taxa μ .

O lado direito da equação (19) corresponde ao fluxo para dentro do estado $B = \{acc\}$. Na equação (19) e na Figura H12, a transição de estado para os vértices adjacentes a $\{acc\}$ são:

- $\{ac\} \rightarrow \{acc\}$, quando ocorre uma chamada nos subátomos $1c$, $2c$, ou $3c$, com taxa λ_c ;
- $\{cc\} \rightarrow \{acc\}$, quando ocorre uma chamada nos subátomos $1a$, $2a$ ou $3a$, com taxa λ_a .

VII. O estado $\{bbc\}$, com seus vértices adjacentes $\{bb\}, \{bc\}$, é mostrado na Figura H13. A equação de equilíbrio do estado $\{bbc\}$ é dada por:

$$\mu P_{\{bbc\}} = \lambda_b P_{\{bc\}} + \lambda_c P_{\{bb\}} \quad (20)$$

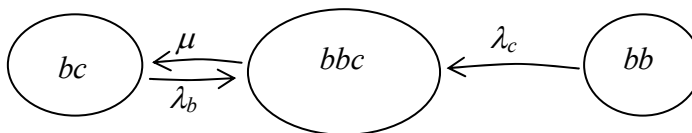


Figura H13 – Vértice $\{bbc\}$ e seus adjacentes.

O lado esquerdo da equação (20) corresponde ao fluxo para fora do estado $B = \{bbc\}$. Na equação (20) e na Figura H13, a transição de estado para o vértice adjacente a $\{bbc\}$ é:

➤ $\{bbc\} \rightarrow \{bc\}$, quando ocorre um término de serviço, com taxa μ .

O lado direito da equação (20) corresponde ao fluxo para dentro do estado $B = \{bbc\}$. Na equação (20) e na Figura H13, a transição de estado para os vértices adjacentes a $\{bbc\}$ são:

➤ $\{bb\} \rightarrow \{bbc\}$, quando ocorre uma chamada nos subátomos $1c$, $2c$, ou $3c$, com taxa λ_c ;

➤ $\{bc\} \rightarrow \{bbc\}$, quando ocorre uma chamada nos subátomos $1b$, $2b$ ou $3b$, com taxa λ_b ;

VIII. O estado $\{bbb\}$, com seu vértice adjacente $\{bb\}$, é mostrado na Figura H14. A equação de equilíbrio do estado $\{bbb\}$ é dada por:

$$\mu P_{\{bbb\}} = \lambda_b P_{\{bb\}} \quad (21)$$

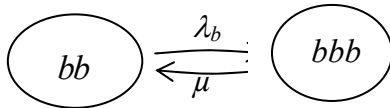


Figura H14 – Vértice $\{bbb\}$ e seu adjacente.

O lado esquerdo da equação (21) corresponde ao fluxo para fora do estado $B = \{bbb\}$. Na equação (21) e na Figura H14, a transição de estado para o vértice adjacente a $\{bbb\}$ é:

➤ $\{bbb\} \rightarrow \{bb\}$, quando ocorre um término de serviço, com taxa μ .

O lado direito da equação (21) corresponde ao fluxo para dentro do estado $B = \{bbb\}$. Na equação (21) e na Figura H14, a transição de estado para o vértice

adjacente a $\{bbb\}$ é:

➤ $\{bb\} \rightarrow \{bbb\}$, quando ocorre uma chamada nos subátomos $1b$, $2b$ ou $3b$, com taxa λ_b

IX. O estado $\{bcc\}$, com seus vértices adjacentes $\{bc\}, \{cc\}$, é mostrado na Figura H15. A equação de equilíbrio do estado $\{bcc\}$ é dada por:

$$\mu P_{\{bcc\}} = \lambda_c P_{\{bc\}} + \lambda_b P_{\{cc\}} \quad (22)$$

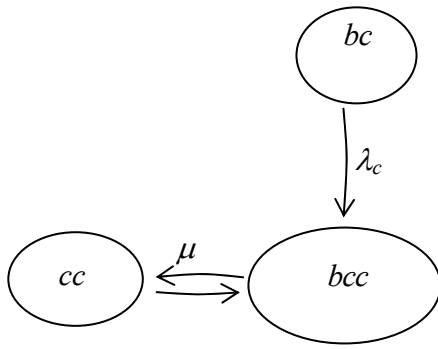


Figura H15 – Vértice $\{bcc\}$ e seus adjacentes.

O lado esquerdo da equação (22) corresponde ao fluxo para fora do estado $B = \{bcc\}$. Na equação (22) e na Figura H15, a transição de estado para o vértice adjacente a $\{bcc\}$ é:

➤ $\{bcc\} \rightarrow \{cc\}$, quando ocorre um término de serviço, com taxa μ .

O lado direito da equação (22) corresponde ao fluxo para dentro do estado $B = \{bcc\}$. Na equação (22) e na Figura H15, a transição de estado para os vértices adjacentes a $\{bcc\}$ são:

➤ $\{bc\} \rightarrow \{bcc\}$, quando ocorre uma chamada nos átomos $1c$, $2c$, ou $3c$, com taxa λ_c ;

➤ $\{cc\} \rightarrow \{bcc\}$, quando ocorre uma chamada nos subátomos $1b$, $2b$ ou $3b$, com taxa λ_b .

X. O estado $\{ccc\}$, com seu vértice adjacente $\{cc\}$, é mostrado na Figura H16.

A equação de equilíbrio do estado $\{ccc\}$ é dada por:

$$\mu P_{\{ccc\}} = \lambda_c P_{\{cc\}} \quad (23)$$

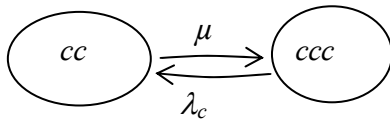


Figura H16 – Vértice $\{ccc\}$ e seus adjacentes.

O lado esquerdo da equação (23) corresponde ao fluxo para fora do estado $B = \{ccc\}$. Na equação (23) e na Figura H16, a transição de estado para o vértice adjacente a $\{ccc\}$ é:

➤ $\{ccc\} \rightarrow \{cc\}$, quando ocorre um término de serviço, com taxa μ .

O lado direito da equação (23) corresponde ao fluxo para dentro do estado $B = \{ccc\}$. Na equação (23) e na Figura H16, a transição de estado para o vértice adjacente a $\{ccc\}$ é:

➤ $\{cc\} \rightarrow \{ccc\}$, quando ocorre uma chamada nos subátomos $1c$, $2c$, ou $3c$, com taxa λ_c ;