**Universidade Federal de Goiás – UFG**

**Escola de Engenharia Elétrica, Mecânica e de Computação**

**Programa de Pós-Graduação em Engenharia Elétrica e de Computação**

Ricardo Bruno Osés de Oliveira

# Aplicação de Algoritmos de Controle e Balanceamento de Carga a um Sistema Perinatal

Goiânia

2021

UNIVERSIDADE FEDERAL DE GOIÁS
ESCOLA DE ENGENHARIA ELÉTRICA, MECÂNICA E DE COMPUTAÇÃO

# TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

# E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

**1. Identificação do material bibliográfico**

[ X ] Dissertação [ ] Tese

**2. Nome completo do autor**

Ricardo Bruno Osés de Oliveira

**3. Título do trabalho**

**“Aplicação de Algoritmos de Controle e Balanceamento de Carga a um Sistema Perinatal”**

**4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)**

Concorda com a liberação total do documento [ X ] SIM [ ] NÃO[1]

**[1]** Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
**a)** consulta ao(à) autor(a) e ao(à) orientador(a);
**b)** novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

**Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.**

Documento assinado eletronicamente por **Flávio Geraldo Coelho Rocha**, **Professor do Magistério Superior**, em 23/03/2021, às 00:13, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **RICARDO BRUNO OSÉS DE OLIVEIRA**, **Discente**, em 23/03/2021, às 15:55, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1958493** e o código CRC **534F80E8**.

**Referência:** Processo nº 23070.014996/2021-45 SEI nº 1958493

Ricardo Bruno Osés de Oliveira

# Aplicação de Algoritmos de Controle e Balanceamento de Carga a um Sistema Perinatal

Dissertação apresentada ao Programa de Pós–Graduação da Escola de Engenharia Elétrica, Mecânica e de Computação (EMC) da Universidade Federal de Goiás, como requisito parcial para obtenção do título de Mestre em Engenharia Elétrica e de Computação.

Universidade Federal de Goiás – UFG

Escola de Engenharia Elétrica, Mecânica e de Computação

Programa de Pós-Graduação em Engenharia Elétrica e de Computação

Área de concentração: Engenharia de Computação

Orientador: Prof. Dr. Flávio Geraldo Coelho Rocha

Goiânia

2021

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Oliveira, Ricardo Bruno Osés de
Aplicação de Algoritmos de Controle e Balanceamento de Carga a um Sistema Perinatal [manuscrito] / Ricardo Bruno Osés de Oliveira. 2021.
C, 100 f.

Orientador: Prof. Dr. Flávio Geraldo Coelho Rocha.
Dissertação (Mestrado) - Universidade Federal de Goiás, Escola de Engenharia Elétrica, Mecânica e de Computação (EMC), Programa de Pós-Graduação em Engenharia Elétrica e de Computação, Goiânia, 2021.
Bibliografia.
Inclui siglas, mapas, abreviaturas, símbolos, gráfico, tabelas, algoritmos, lista de figuras, lista de tabelas.

1. Balanceamento de Carga. 2. Markov. 3. Perinatal. 4. Maternidade. 5. Comportamento das Abelhas. I. Rocha, Flávio Geraldo Coelho, orient. II. Título.

CDU 621.3

UNIVERSIDADE FEDERAL DE GOIÁS

ESCOLA DE ENGENHARIA ELÉTRICA, MECÂNICA E DE COMPUTAÇÃO

**ATA DE DEFESA DE DISSERTAÇÃO**

Ata nº 01 da sessão de Defesa de Dissertação de Ricardo Bruno Osés de Oliveira, que confere o título de Mestre em **Engenharia Elétrica e de Computação**, na área de concentração em **Engenharia de Computação**.

Aos **dezoito dias do mês de fevereiro de dois mil e vinte e um**, a partir das **14h30min.**, realizou-se a sessão pública de Defesa de Dissertação intitulada **"Aplicação de Algoritmos de Controle e Balanceamento de Carga a um Sistema Perinatal"**. Os trabalhos foram instalados pelo Orientador, Professor Doutor **Flávio Geraldo Coelho Rocha (EMC/UFG),** com a participação dos demais membros da Banca Examinadora: Professor Doutor **Leizer de Lima Pinto (INF/UFG)**, membro titular externo; Professor Doutor **Marcelo Stehling de Castro (EMC/UFG)**, membro titular externo; Professor Doutor **Rodrigo Pinto Lemos (EMC/UFG)**, membro titular interno. Durante a arguição os membros da banca **não fizeram** sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação, tendo sido o candidato **aprovado** pelos seus membros. Proclamados os resultados pelo Professor Doutor **Flávio Geraldo Coelho Rocha**, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos **dezoito dias do mês de fevereiro de dois mil e vinte e um**.

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Flávio Geraldo Coelho Rocha**, **Professor do Magistério Superior**, em 18/02/2021, às 16:45, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Rodrigo Pinto Lemos**, **Professor do Magistério Superior**, em 18/02/2021, às 16:51, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Marcelo Stehling De Castro**, **Professor do Magistério Superior**, em 18/02/2021, às 20:53, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Leizer De Lima Pinto**, **Professor do Magistério Superior**, em 18/02/2021, às 22:38, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **RICARDO BRUNO OSÉS DE OLIVEIRA**, **Discente**, em 20/02/2021, às 15:21, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1878857** e o código CRC **8DE5918B**.

**Referência:** Processo nº 23070.005068/2021-90 SEI nº 1878857

*Ao meu irmão, Augusto Sérgio de Oliveira.*

# Agradecimentos

Gostaria de expressar meus agradecimentos

A Deus, por esta oportunidade.

Aos meus pais, Geraldo e Rivany, pelo apoio, incentivo, por sempre acreditarem em meus sonhos e por nunca medirem esforços para investir na minha educação.

À minha esposa, Ana Carolina, pelo carinho, apoio e compreensão durante esta jornada.

Aos meu amigos de Goianésia-GO e aos amigos que fiz no INCOMM, pela ajuda, apoio e incentivo.

Ao meu orientador, Prof. Dr. Flávio Geraldo Coelho Rocha, que sempre acreditou em mim.

Aos membros da banca pelas valiosas sugestões e comentários.

À CAPES pelo apoio financeiro.

*"Não vos amoldeis às estruturas deste mundo,*
*mas transformai-vos pela renovação da mente,*
*a fim de distinguir qual é a vontade de Deus:*
*o que é bom, o que Lhe é agradável, o que é perfeito.*
*(Bíblia Sagrada, Romanos 12, 2)*

# Resumo

Nesta dissertação, é considerado o processo de admissão e escalonamento de gestantes em uma rede perinatal composta por duas maternidades. São descritos os principais setores de serviços que compõem cada unidade perinatal e suas respectivas funções e recursos utilizados. Além disso, é feito um levantamento dos principais desafios e dificuldades enfrentados por essas unidades de saúde no Brasil nas últimas décadas. Assim, diante dos problemas e desafios apontados, são utilizados diferentes algoritmos de balanceamento de carga na rede perinatal proposta, a fim de encontrar a melhor política de escalonamento de tarefas no sistema que aumente a eficiência da rede. É proposta uma solução para um problema de programação linear inteira mista que utiliza um algoritmo de balanceamento de carga baseado na meta-heurística do comportamento das abelhas produtoras de mel. Além disso, são analisados alguns algoritmos onde cada um é diferenciado por uma estratégia de roteamento de tarefas projetada para reduzir o tempo médio de atendimento às gestantes que entram no sistema perinatal, equilibrando a carga de trabalho entre as maternidades. São utilizadas duas classes de roteamento, não determinística e determinística. Na classe determinística, são analisadas três políticas de roteamento que buscam diminuir o tempo médio de permanência, o tempo médio de atendimento ou melhorar a vazão no sistema. Adicionalmente, é analisada também, uma política de controle dinâmico com enfileiramento baseado em um limiar, onde um comprimento específico da fila é definido e identificado por um limite. Também é analisada a política de roteamento de junção a fila mais curta, onde cada gestante que entra no sistema é encaminhada para a maternidade com a menor fila de espera. Os resultados são apresentados e analisados variando tanto as taxas de chegada das gestantes, quanto as taxas de atendimentos nos principais setores existentes em uma maternidade. Além disso, é feito um modelo de simulação de eventos discretos para analisar o tempo de espera nas filas. Por fim, utilizando-se da fórmula de *Erlang-B*, é feito o cálculo da capacidade das unidades perinatais com base no tempo de permanência das gestantes no sistema obtido por meio dos algoritmos de balanceamento de carga. Os resultados obtidos confirmam que, políticas de roteamento e escalonamento que levam em consideração a taxa de chegadas de tarefas e o comprimento de fila do sistema são mais eficientes à medida que a taxa de chegadas aumenta, sendo, portanto aplicáveis em sistemas de saúde com uma crescente demanda e que com planejamento é possível obter uma descrição precisa do número de leitos ocupados e do número de leitos necessários de acordo com a demanda exigida pelas unidades perinatais tornando essas unidades mais eficientes.

**Palavras-chave:** Balanceamento de Carga, Markov, Perinatal, Maternidade, Comportamento das Abelhas.

# Abstract

In this work, the process of admission and scheduling of pregnant women in a perinatal network composed of two maternities is considered. The main service sectors that make up each perinatal unit and their respective functions and resources used are described. In addition, a survey is made of the main challenges and difficulties faced by these health units in Brazil in recent decades. Thus, given the problems and challenges pointed out, different load balancing algorithms are used in the proposed perinatal network, to find the best task scheduling policy in the system that increases the efficiency of the network. A solution for a mixed integer linear programming problem is proposed using a load balancing algorithm based on metaheuristics of the behavior of honey bees. Besides, some algorithms are analyzed where each considered model has its own task routing strategy designed to reduce the average time of the pregnant women entering the perinatal system, balancing the workload among the perinatal care centers. Two classes of routing are used, non-deterministic and deterministic. In the deterministic class, three routing policies are analyzed that seek to decrease the average stay time, the average service time, or to improve the flow in the system. In addition, a dynamic control policy based on a queuing threshold is also analyzed, where a specific queue length is defined and identified by a threshold. The routing policy named Join-the-Shortest-Queue (JSQ) is also analyzed, where each pregnant woman who enters the system is directed to the maternity ward with the shortest queue. The results are presented and analyzed varying both the arrival rates of pregnant women and the rates of care in the main sectors in a maternity hospital. Also, a discrete event simulation model is made to analyze the waiting time in queues. Finally, using the formula of *Erlang-B*, the capacity of the perinatal units is calculated based on the time of permanence of the pregnant women in the system obtained through the load balancing algorithms. The results obtained confirm that routing and scheduling policies that consider the task arrival rate and the system queue length are more efficient as the arrival rate increases, therefore being applicable in healthcare systems with increasing demand and that with planning it is possible to obtain an accurate description of the number of occupied beds and the number of beds needed according to the demand required by the perinatal units making these units more efficient.

**Keywords:** Load Balancing, Markov, Perinatal, Maternity, Honey Bee Behavior.

# Lista de ilustrações

# Lista de tabelas

# Lista de abreviaturas e siglas

| | |
|---|---|
| ABC | *Artificial Bee Colony Algorithm* |
| ACO | *Ant Colony Optimization* |
| BA | *Bees Algorithm* |
| BCPA | *Bee Collecting Pollen Algorithm* |
| BS | *Bee System* |
| CC | *Centro Cirúrgico obstétrico* |
| CNES | *Cadastro Nacional de Estabelecimentos de Saúde* |
| CO | *Centro Obstétrico* |
| EBG | *Equação de Balanço Global* |
| FCFS | *First-Come-First-Served* |
| HBA | *Honey Bee Algorithm* |
| HBB-LB | *Honey Bee Behavior Inspired Load Balancing* |
| HBMO | *Honey-Bee Mating Optimization* |
| HOL | *Head-of-Line* |
| IBGE | *Instituto Brasileiro de Geografia e Estatística* |
| IEEE | *Institute of Electrical and Electronic Engineers* |
| JSQ | *Join the Shortest Queue* |
| LBC | *Load-Balancing Policy with a Central Job Dispatcher* |
| MBO | *Mating Bee Optimization* |
| MDP | *Markov Decision Process* |
| NV | *Nascidos Vivos* |
| OMS | *Organização Mundial da Saúde* |
| PSO | *Particle Swarm Optimization* |

| | |
|---|---|
| QEGA | *Queen-Bee Evolution for Genetic Algorithms* |
| RPR | *Resilient Packet Ring* |
| SUS | *Sistema Único de Saúde* |
| TMI | *Taxa de Mortalidade Infantil* |
| TT | *Turnaround Time* |
| UTI | *Unidade de Terapia Intensiva* |
| VBA | *Virtual Bee Algorithm* |

# Lista de símbolos

| | |
|---|---|
| A | *Matriz tridiagonal com os valores das taxas de transições entre os estados* |
| $a_{ij}$ | *Variável de decisão que verifica se gestante está admitida na maternidade* |
| $b_{ij}$ | *Variável de decisão que verifica se gestante está na fila de espera da maternidade* |
| C | *Função de critério utilizada pelas estratégias de roteamento* |
| $c_o$ | *Índice que representa as colunas da matriz de probabilidades* |
| CT | *Tempo de atendimento* |
| EXP | *Função exponencial* |
| K | *Número médio de tarefas na fila* |
| L | *Limite do sistema* |
| $l_i$ | *Índice que representa as linhas da matriz de probabilidades* |
| $\lambda_i$ | *Taxa de chegadas de gestantes* |
| $\mu_i$ | *Taxa de serviços das maternidades* |
| M | *Conjunto de maternidades* |
| m | *Número de servidores* |
| $M_o$ | *Maternidades com sobrecarga* |
| $M_u$ | *Maternidades com baixa carga* |
| N | *Número de filas de espera* |
| $n_i$ | *Número de tarefas na i-ésima fila* |
| $P$ | *Conjunto de gestantes* |
| $\pi$ | *Probabilidades de estado em regime permanente* |
| $p(s)$ | *Probabilidade de estado estacionário do estado s* |
| $P_{ij}$ | *Matriz de probabilidades de estado estacionário* |
| $PR$ | *Algoritmo não determinístico* |

| | |
|---|---|
| $\rho$ | *Intensidade do tráfego ou o fator de utilização* |
| $Q$ | *Conjunto de filas* |
| $QP$ | *Quantidade total de gestantes atendidas* |
| $R$ | *Política de Tempo Mínimo de Resposta* |
| s | *Estado do sistema* |
| $T$ | *Política de Tempo Mínimo de Sistema* |
| $TP$ | *Política de Máxima Vazão* |
| $x$ | *Vetor de incógnitas* |
| $WQ$ | *Número de gestantes na fila de espera* |
| $WT$ | *Tempo de espera* |
| $y$ | *Vetor com valores obtidos através da recursão* |

# Trabalhos Publicados

1. OLIVEIRA, R. B. O.; ROCHA, F. G. C. . Balanceamento Dinâmico de Carga em um Sistema Heterogêneo com dois Servidores. Em: *X Conferência Nacional em Comunicação, Redes e Segurança da Informação* (ENCOM), Natal - RN , 2020.

2. LOPES, H. H. S. ; OLIVEIRA, R. B. O. ; ROCHA, F. G. C. . Balanceamento de Carga Inspirado no Comportamento das Abelhas Aplicado a uma Rede Perinatal. Em: *XIX Workshop em Desempenho de Sistemas Computacionais e de Comunicação* (WPERFORMANCE), 2020, Cuiabá - MT. Anais do XIX Workshop em Desempenho de Sistemas Computacionais e de Comunicação, 2020.

3. OLIVEIRA, R. B. O.; ROCHA, F. G. C. . Controle Dinâmico de Carga Baseado em Política de Limiar Aplicado a um Sistema Perinatal. Em: *LII Simpósio Brasileiro de Pesquisa Operacional* (SBPO), 2020, João Pessoa - PB. LII Simpósio Brasileiro de Pesquisa Operacional, 2020.

4. OLIVEIRA, R. B. O.; ROCHA, F. G. C. ; SILVA, M. R. P. . Avaliação de Modelos de Balanceamento de Carga Aplicados a um Sistema Perinatal. Em: *LI Simpósio Brasileiro de Pesquisa Operacional* (SBPO), 2019, Limeira - SP. LI Simpósio Brasileiro de Pesquisa Operacional, 2019.

5. SILVA, M. R. P. ; ROCHA, F. G. C. ; OLIVEIRA, R. B. O. . Modelagem Multifractal para o Tráfego de Dados com Aplicação em um Cenário Simplificado de Rede 5G. Em: *LI Simpósio Brasileiro de Pesquisa Operacional* (SBPO), 2019, Limeira - SP. LI Simpósio Brasileiro de Pesquisa Operacional, 2019.

# Sumário

# 1 Introdução

O sistema de saúde representa uma parte muito importante do setor de serviços de um país. Em particular, hospitais de rede pública e privada desempenham um papel muito importante em qualquer sociedade. Ao longo dos anos, os hospitais obtiveram sucesso no uso de inovações médicas e técnicas para oferecerem tratamentos clínicos mais eficazes e reduzirem o tempo gasto pelos pacientes no hospital. No entanto, os hospitais ainda apresentam atrasos de atendimento e ineficiências em geral, sendo, portanto, um terreno propício para muitos projetos de pesquisa em vários campos da ciência (TSEYTLIN; ZVIRAN, 2008).

A maioria dos hospitais possui diversas unidades médicas especializadas em diferentes áreas da medicina operando em paralelo, como, unidades de terapia intensiva, cirurgia, perinatal, dentre outras. Nesta pesquisa, os estudos são concentrados nas unidades de saúde perinatal. Por perinatal, entende-se tudo o que acontece entre a 22ª semana de gestação e a primeira semana de vida do recém-nascido. As unidades de saúde perinatal no Brasil são responsáveis por esses períodos próximos ao nascimento e enfrentam múltiplos desafios atualmente: o aumento no número de gestações múltiplas, a diminuição do financiamento do governo na área da saúde e a falta de profissionais capacitados em unidades públicas, onde ocorre a maior parte dos atendimentos perinatais (LEAL et al., 2015), (PEREIRA et al., 2018). Outro desafio dessas unidades de saúde é oferecer acesso aos serviços obstétricos de forma igualitária (LEAL et al., 2015). Há desigualdade na distribuição de unidades, leitos e uma estrutura adequada às maternidades (CAMPOS; CARVALHO, 2000), (BITTENCOURT et al., 2014), (ALVES et al., 2014), (SILVA et al., 2017).

Além da desigualdade na disponibilidade de serviços e recursos de saúde, problemas de acesso geográfico aos centros de saúde com infraestrutura de suporte à gestante em trabalho de parto refletem as falhas na integração e articulação entre os setores da saúde. Apesar de o parto possuir um tempo previsto para ocorrer, a atenção materno-infantil mantém-se desarticulada e fragmentada (COSTA et al., 2009), (ALMEIDA; SZWARCWALD, 2012). Se o acesso aos serviços de saúde é fundamental para uma assistência eficiente e para a redução das desigualdades, o enfoque à inacessibilidade de alguns grupos populacionais é essencial para a tomada de decisão sobre a localização e dimensão dos serviços. O acesso geográfico a estes centros de saúde é, portanto, fundamental para qualificar a assistência materno-infantil (UNGLERT, 1990), (ALMEIDA; SZWARCWALD, 2012).

A alocação dos recursos das maternidades também se torna algo desafiador, já que, determinados tipos de serviços oferecidos pela rede perinatal utilizam uma quantidade

considerável de recursos financeiros, mesmo se não efetivamente utilizados. Leitos ociosos, equipamentos ou recursos humanos não utilizados em uma maternidade podem significar a falta desses mesmos recursos em outra maternidade próxima. Como consequência, elevar o investimento público e privado de maneira indiscriminada em unidades de atendimento perinatais não significa necessariamente elevar a quantidade e qualidade dos atendimentos oferecidos à comunidade. Tanto o investimento quanto a operação eficiente em unidades de saúde perinatais são vitais, visto que a demora ou não atendimento das gestantes podem causar possíveis ocorrências fatais, complicações no parto e impactos diretos na saúde das mães e dos recém-nascidos (RêGO et al., 2018).

As pacientes de redes perinatais são urgentes, não podem esperar muito. Portanto, a capacidade de atendimento inadequada em uma maternidade faz com que a gestante espere muito ou seja transferida para outra instituição, o que pode causar longas distâncias de deslocamento para uma unidade de saúde não preferida. Em outras palavras, a capacidade inadequada em uma unidade resulta transbordamento da fila de gestantes e, consequentemente, rejeição de atendimentos, o que pode e deve ser evitado por meio de um apropriado dimensionamento e operação eficiente dos recursos disponíveis (PEHLIVAN; AUGUSTO; XIE, 2013). Assim, é essencial que a correta utilização dos recursos e a qualidade do serviço prestado por unidades perinatais sejam eficazes, diminuindo o tempo de atendimento, a permanência das gestantes na rede e o número de ocorrências fatais.

Diante dos desafios apresentados pelo sistema de saúde em geral, muitos projetos aplicados à saúde têm sido desenvolvidos. Green (GREEN, 2005), (GREEN, 2008) descreve os antecedentes gerais e as questões envolvidas no planejamento da capacidade hospitalar e mostra exemplos de como as metodologias de Pesquisa Operacional podem ser usadas para fornecer informações importantes sobre estratégias e práticas operacionais. Outro trabalho apresentado por (MESTRE; OLIVEIRA; BARBOSA-PÓVOA, 2012) propõe um modelo hierárquico de programação matemática multisserviço para informar decisões sobre a localização e a prestação de serviços hospitalares, quando o tomador de decisão deseja maximizar o acesso geográfico dos pacientes a uma rede hospitalar.

Tseytlin e Zviran (TSEYTLIN; ZVIRAN, 2008) consideraram o processo de encaminhamento dos pacientes de um departamento de emergência para enfermarias internas em um hospital anônimo. Buscando diminuir o tempo de espera dos pacientes que aguardam uma transferência do pronto-socorro para as enfermarias e garantir uma alocação justa destes pacientes, os autores modelaram o processo de transferência como um sistema de enfileiramento com *pools* de servidores heterogêneos, onde os *pools* representam as enfermarias e os servidores são camas. Esse sistema foi analisado sob várias arquiteturas e políticas de roteamento, em busca de atender a requisitos de índices de justiça e desempenho operacional.

Pehlivan (PEHLIVAN, 2014), motivada pelos desafios das redes perinatais, abordou

o *design* e o controle de fluxo de uma rede de saúde estocástica, onde existem vários níveis de hospitais, recursos e diferentes tipos de pacientes. De acordo com o problema de perspectivas estratégicas e operacionais foram desenvolvidas metodologias de ponta para avaliar o desempenho de um sistema complexo e melhorar a eficiência, diminuindo as taxas de rejeição na rede, com o objetivo final de ser aplicado à rede perinatal real da França.

Ramakrishnan, Sier e Taylor (RAMAKRISHNAN; SIER; TAYLOR, 2005) construíram um modelo em duas escalas para um sistema hospitalar, onde as enfermarias operam em uma escala de dias e são modeladas por uma cadeia Markov de tempo discreto, e o departamento de emergência opera em uma escala de tempo dada em horas e é modelado por uma cadeia de Markov de tempo contínuo. Com a ajuda desse modelo, os autores estimaram a ocupação esperada das enfermarias e a probabilidade de cada ala atingir sua capacidade.

Augusto, Murgier e Viallon (AUGUSTO; MURGIER; VIALLON, 2018) propuseram uma estrutura de modelagem e simulação para a avaliação de desempenho de unidades de emergência e controle inteligente em caso de grande crise. É proposta uma ferramenta flexível que pode ser usada pelos profissionais de saúde como auxílio à decisão em várias situações sob a forma de uma representação digital da unidade de emergência.

Prodel, Augusto e Xie (PRODEL; AUGUSTO; XIE, 2014) abordaram as políticas de controle de internação de pacientes de um serviço de emergência que devem ser admitidos ou transferidos imediatamente. O problema consiste em determinar as políticas de admissão de pacientes para aumentar o ganho geral do hospital. Primeiro, foi proposto um modelo de Processo de Decisão de Markov (MDP) para determinar a política ideal de admissão de pacientes sob algumas premissas restritivas e necessárias. Um modelo de simulação foi então construído para avaliar as políticas de admissão do MDP sob condições realistas e os resultados mostraram que as políticas do MDP melhoram significativamente o ganho geral para diferentes tipos de instalações.

Nesse contexto, os diversos avanços científicos e tecnológicos possibilitaram a aplicação do conhecimento desenvolvido em áreas como a engenharia, a matemática e a Pesquisa Operacional na resolução de problemas enfrentados pelos sistemas de saúde. De fato, os trabalhos supracitados propõem métodos matemáticos, modelos e ferramentas que são elegíveis para serem utilizados em cenários reais de sistemas de saúde. Alguns dos objetivos desses trabalhos são: melhorar a qualidade dos serviços prestados pelas unidades hospitalares, diminuindo o tempo de espera na utilização dos serviços oferecidos, distribuir geograficamente, de maneira ótima, essas unidades de forma a atenderem um maior número de pacientes, diminuir os custos de utilização dos recursos disponíveis, prever eventos de crise em que há grande demanda por serviços de urgência, dentre outros.

Com base no que foi exposto, a partir dos diversos projetos de pesquisa na área da saúde em geral e, mais especificamente, analisando o cenário desigual de distribuição de

recursos entre unidades de saúde perinatal no Brasil, motivou-se investigar o impacto da aplicação de algoritmos de balanceamento dinâmico de carga em um sistema perinatal composto por múltiplas unidades de atendimento com capacidades heterogêneas. É feito um estudo de caso para duas maternidades buscando encontrar a melhor política de atendimento e encaminhamento de gestantes que diminuam o tempo de espera nessas unidades fazendo uma melhor utilização dos recursos disponíveis.

Esta dissertação propõe:

1. Um algoritmo de balanceamento de carga baseado na meta-heurística do comportamento das abelhas melíferas para alocação eficiente de recursos no processo de admissão e agendamento de gestantes em uma rede perinatal composta por duas maternidades com capacidades heterogêneas;

2. Aplicação de algoritmos de balanceamento de carga encontrados na literatura com o objetivo de validar os resultados obtidos pelo algoritmo proposto;

3. Simulação de eventos discretos para o sistema perinatal analisado, apresentando os tempos médios de espera nas filas das maternidades;

4. Dimensionamento dos recursos das maternidades com base no tempo de permanência das gestantes no sistema obtidos pelos algoritmos de balanceamento de carga. Para tal finalidade, utiliza-se a fórmula de *Erlang-B* para o cálculo e estimativa da quantidade de leitos das unidades perinatais com o objetivo de diminuir as taxas de rejeição.

O restante desta dissertação está organizado da seguinte forma: no Capítulo 2 são apresentadas informações sobre as políticas de controle e admissão à rede perinatal. Descreve-se a rede perinatal utilizada e a abordagem sistêmica de uma maternidade; no Capítulo 3 introduz-se as bases teóricas dessa pesquisa. São apresentados alguns algoritmos de balanceamento e controle de carga presentes na literatura alem da proposta de um algoritmo de balanceamento de carga inspirado no comportamento das abelhas produtoras de mel; no Capítulo 4 é apresentado o modelo de simulação para um sistema perinatal composto por duas maternidades com capacidades heterogêneas. Os resultados obtidos pelos algoritmos de balanceamento de carga para o modelo são apresentados, analisados e comparados; no Capítulo 5 é introduzido o modelo de perda de Erlang, onde este é utilizado para o dimensionamento da capacidade das duas maternidades analisadas; por fim, no Capítulo 6 são apresentadas as conclusões obtidas.

# 2 Políticas de Controle e Admissão à Rede Perinatal

Nesta seção é analisado e descrito o fluxo de gestantes entre duas maternidades. São fornecidas informações sobre a rede perinatal estudada (Seção 2.1), todo o processo de admissão e encaminhamento das gestantes (Seção 2.2) e os principais problemas envolvidos (Seção 2.3) que motivaram este estudo.

## 2.1 Rede Perinatal

A rede perinatal considerada neste estudo está representada na Figura 2.1. Têm-se duas classes de gestantes que chegam à rede. A classe 1 é de gestantes que são admitidas na primeira maternidade em que foram atendidas sem a necessidade de transferência para outra maternidade, já a classe 2 é de gestantes que foram transferidas de uma maternidade para outra. Se as gestantes são rejeitadas por uma das maternidades ($M_1$ ou $M_2$), é feita a transferência dessas gestantes entre as maternidades.

Assume-se que as chegadas das gestantes nas maternidades podem ser descritas por meio de processos de Poisson (BRANDEAU; SAINFORT; PIERSKALLA, 2004). Portanto, presume-se que as gestantes chegam a uma unidade de serviço na maternidade $M_i$, onde $i = \{1, 2\}$, com periodicidade dada segundo uma distribuição de Poisson com parâmetro $\lambda_i$ e os atendimentos ocorrem com uma taxa média de serviço exponencial com parâmetro $\mu_i$.

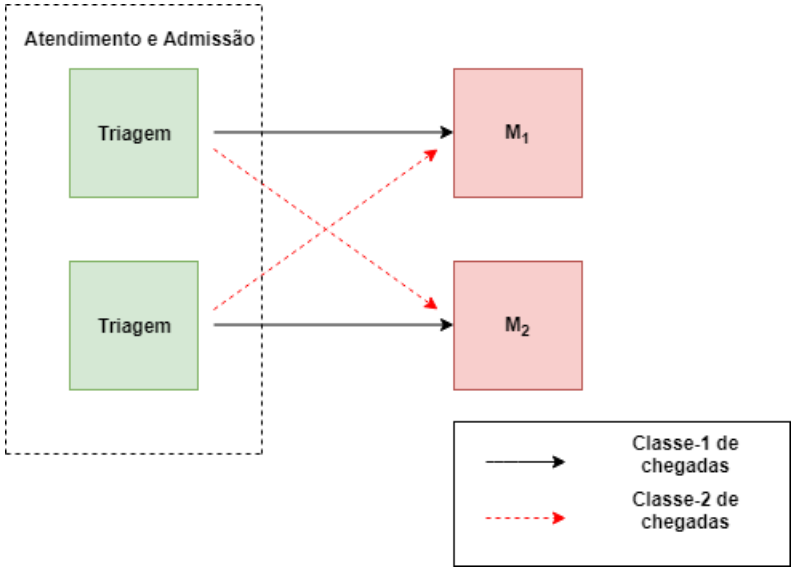



Figura 2.1 – Representação da Rede Perinatal.

## 2.2 O Sistema Maternidade

A abordagem sistêmica dada à unidade perinatal revela a interação existente entre seus subsistemas. Têm-se os setores de obstetrícia, cirurgia e Unidade de Terapia Intensiva (UTI) como setores essenciais da cadeia de serviços da maternidade, inteirando-se diretamente com os leitos, salas de recuperação e outras áreas de serviços hospitalares. Em algumas maternidades os setores de cirurgia são contínuos aos setores de obstetrícia, enquanto em outras, estes setores funcionam em alas separadas. Para esta pesquisa, considerou-se que, estes setores funcionam como unidades separadas dentro das maternidades. Os Centros Obstétricos (CO) e Cirúrgicos obstétricos (CC) são as áreas de maior serviço em uma maternidade, e exercem importante influência no fluxo de gestantes. Além disso, seus serviços implicam em altos custos, já que devem possuir um quadro de profissionais altamente especializados e necessitam de grandes investimentos de capital em aparelhagem e infraestrutura. À luz destes altos custos operacionais e da possibilidade destes centros se tornarem subutilizados ou congestionados, é obrigatório que os mesmos funcionem com um alto nível de eficiência.

### 2.2.1 Centros Obstétricos

Os Centros Obstétricos (CO) constituem-se em unidades de atendimento ao parto normal localizadas fora do Centro Cirúrgico obstétrico (CC). Dispõe de um conjunto de elementos destinados a receber a parturiente e seu acompanhante, de forma humanizada e que permita a evolução do parto o mais fisiológico possível, ativo, participativo e, sobretudo seguro. São compostos por salas com leitos de obstetrícia clínica para pré-parto, parto e pós-parto. As salas com leitos para pré-parto são destinadas ao recebimento e acompanhamento de parturientes nos momentos que antecedem o parto. Nestas unidades, é realizado o constante monitoramento materno fetal com objetivo de avaliar o progresso do trabalho de parto, garantindo a segurança da parturiente e do feto. As salas com leitos para parto são as unidades de atendimento de realização do parto normal.

### 2.2.2 Centros Cirúrgicos Obstétricos

Os Centros Cirúrgicos obstétricos (CC) constituem-se em um conjunto de áreas e instalações destinadas ao atendimento ao parto cesáreo. São compostos por salas de cirurgia e unidades de atendimento pós-anestesia. Geralmente atendem a partos de emergência, que contenham certo risco de vida a parturiente ou ao feto. Partos cesáreos eletivos também são realizados nessas unidades.

### 2.2.3 Unidades de Terapia Intensiva Neonatal

As Unidades de Terapia Intensiva Neonatal (UTI) fornecem cuidados intensivos para recém-nascidos que apresentem algum tipo de problema ao nascer. O tratamento intensivo é, na maioria das vezes, indicado para recém-nascidos prematuros, ou de baixo peso, com menos de 2,5 Kg. Entretanto, qualquer recém-nascido pode precisar da UTI Neonatal. Grande parte dos casos é de recém-nascidos com dificuldade respiratória. Problemas cardíacos, icterícia ou cirurgias também podem exigir cuidados intensivos. As UTIs têm capacidade para fornecer ventilação mecânica e suporte respiratório avançado, acesso a uma gama completa de subespecialidades médicas dentro da pediatria, de exames de imagens avançadas (incluindo tomografia computadorizada, ressonância magnética e ecocardiografia), especialistas em cirurgia pediátrica e anestesiologistas pediátricos.

### 2.2.4 Encaminhamento das Gestantes

As unidades perinatais oferecem serviços diversos às gestantes que vão desde exames pré-natais, acompanhamento do período de gestação, partos dos tipos normal, cesáreo entre outros. Os partos normais são os mais realizados e, em condições favoráveis em relação a saúde da gestante e do feto, os mais recomendáveis. Os partos cesáreos são geralmente realizados apenas em casos de emergência, quando a mãe ou o bebê estão correndo algum risco de vida, ou quando se trata de partos eletivos, que ocorrem em sua maioria em redes perinatais privadas.

Quanto ao funcionamento do sistema, quando uma gestante chega à maternidade ela passa por um processo em que são verificadas as condições relacionadas ao parto. Se ainda não estiver no momento dessa gestante entrar para a sala de parto, ela é levada para um leito pré-parto, onde fica aguardando um período, antes do ingresso a um dos centros CC ou CO. A gestante permanece ocupando o leito pré-parto até obter o grau de dilatação necessário nos casos obstétricos, ou aguardando a hora, nos casos de cirurgias eletivas. Se, eventualmente, não houver sala de pré-parto disponível, a gestante pode permanecer na espera até que exista possibilidade de sua admissão ao CO.

Na sala de pré-parto o departamento médico determina com base na avaliação do quadro clínico da gestante se o parto será cirúrgico ou não cirúrgico. Os partos cirúrgicos do tipo cesariana são realizados no CC, os partos não cirúrgicos são do tipo normal e são realizados no CO. Após o parto, processa-se a saída da paciente do CC ou CO e seu encaminhamento para o leito pós-parto, onde permanecerá para restabelecimento final, até receber alta. Em casos em que há complicações no parto com risco de vida para a gestante e/ou ao recém-nascido estes podem ser deslocados para a UTI, onde podem permanecer até que se recuperem totalmente.

O diagrama simplificado que modela o sistema de maternidades definido neste

trabalho é representado na Figura 2.2. Foi considerado apenas o fluxo das gestantes, antes e após o parto.

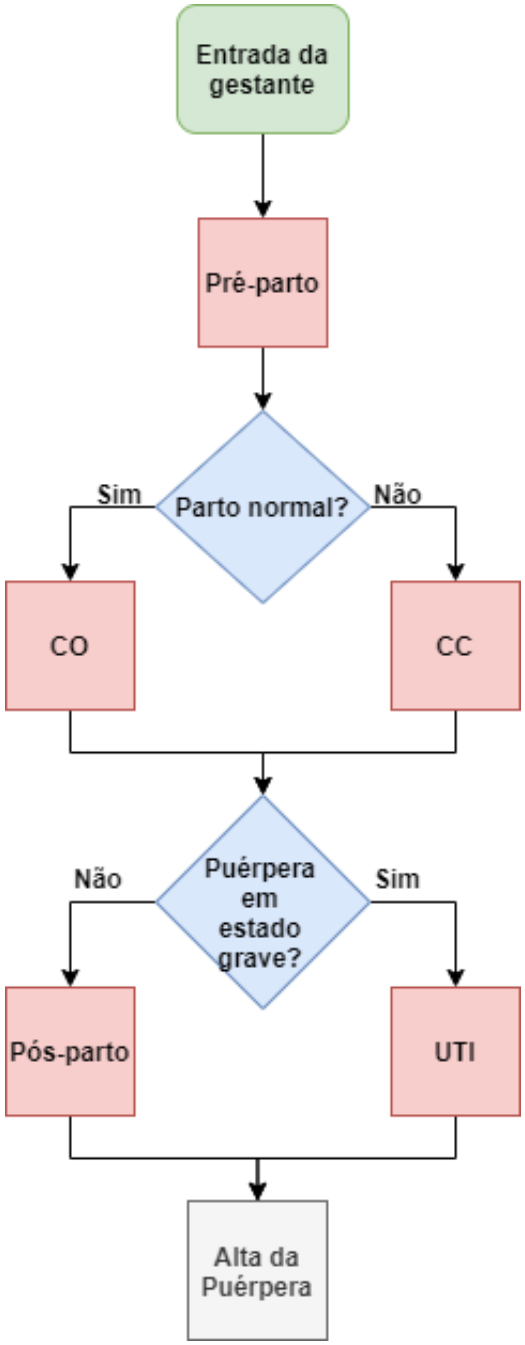



Figura 2.2 – Sistema de uma Maternidade.

## 2.3 Dificuldades e Desafios das Unidades Perinatais no Brasil

Nessa seção são apresentados as principais dificuldades e os desafios enfrentados pelas unidades perinatais no Brasil. São apresentados estatísticas sobre a capacidade instalada e localização das principais unidades de assistência ao parto em todas as regiões

do país. Além disso, é apresentado o comportamento populacional nos últimos anos descrito pelas taxas de natalidade e mortalidade.

### 2.3.1 Redução do Número de Leitos Obstétricos

O número de maternidades no Brasil foi reduzido regularmente na última década segundo levantamento do (SAúDE, 2019). Essa tendência está associada a vários fatores. Um dos principais motivos de fechamento está relacionado ao alto custo gerado por essas unidades, à baixa remuneração e a pouca rentabilidade de um leito de maternidade em comparação com um leito de outra especialidade que envolve materiais e medicações de alto custo. Caso a instituição não tenha um volume adequado de partos, a maternidade se torna inviável financeiramente.

Outro fator provocador é a escassez de recursos, um problema comum na maioria das unidades de saúde, como número insuficiente de ginecologistas, obstetras, enfermeiras obstetras e níveis inadequados de equipamentos. Além disso, políticas governamentais que visam reduzir os custos da prestação de serviços de saúde podem ser apontadas como outro motivo para o fechamento. A Tabela 2.1 (SAúDE, 2019) apresenta o número de leitos obstétricos existentes no Brasil durante o período 2010-2019.

Tabela 2.1 – Número de leitos obstétricos no Brasil 2010-2019.

| **Tipo do leito** | **2010** | **2011** | **2012** | **2013** | **2014** | **2015** | **2016** | **2017** | **2018** | **2019** |
|---|---|---|---|---|---|---|---|---|---|---|
| Obstetrícia clínica | 28598 | 28026 | 27731 | 27068 | 26844 | 26279 | 26117 | 25692 | 25402 | 25151 |
| Obstetrícia cirúrgica | 32021 | 30528 | 30592 | 29979 | 29558 | 28342 | 27841 | 27786 | 27208 | 26746 |
| Total | 60619 | 58554 | 58423 | 57047 | 56402 | 54621 | 53958 | 53478 | 52610 | 51897 |

Na Figura 2.3, construída a partir dos dados da tabela 2.1, é possível observar que houve uma diminuição de cerca de 14,4% no número de leitos obstétricos em um período de 10 anos. O maior risco associado a essa diminuição do número de leitos é a redução da acessibilidade dos serviços oferecidos pelas maternidades, que pode atingir em maior parte as mulheres de classes sociais mais baixas e de regiões mais pobres. Além disso, como em todas as especialidades médicas que envolvem o gerenciamento de emergências hospitalares, o acesso rápido é crucial também nos cuidados perinatais e neonatais, a fim de evitar partos acidentais fora das maternidades e o risco de morbidade materna e neonatal (GULLIFORD et al., 2002).

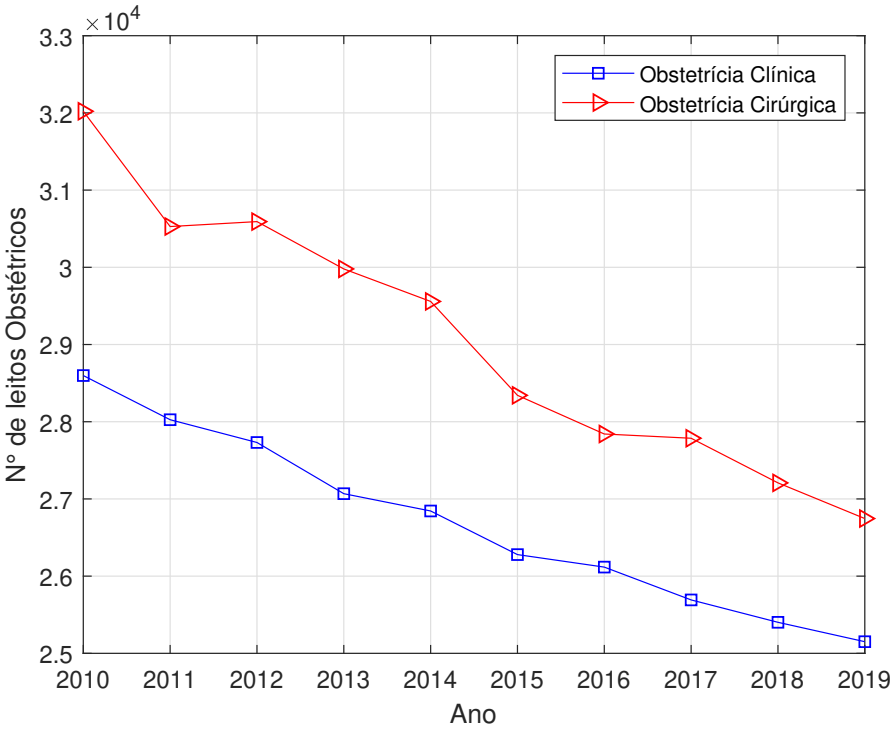



Figura 2.3 – Número de leitos obstétricos no Brasil 2010-2019.

As estatísticas sobre o número de nascidos vivos no Brasil (IBGE, 2018) durante o período de 2010-2018 apresentadas na Tabela 2.2, mostram que há uma evolução crescente no número de partos realizados.

Tabela 2.2 – Número de nascidos vivos no Brasil 2010-2019.

| **2010** | **2011** | **2012** | **2013** | **2014** | **2015** | **2016** | **2017** | **2018** |
|---|---|---|---|---|---|---|---|---|
| 2760961 | 2824514 | 2830458 | 2832590 | 2913121 | 2952969 | 2803080 | 2874466 | 2899851 |

Observa-se na Figura 2.4, construída a partir dos dados da tabela 2.2, um aumento de cerca de 5% no número total de partos, isso desconsiderando o número de nascidos mortos em todo país. Embora tenha ocorrido uma queda de 5,1% em 2016, os registros ainda mostram um aumento do número de nascimentos nos anos seguintes. Entre os anos de 2015 e 2016, o Brasil registrou uma alteração no padrão de ocorrência de microcefalia, que foi associado à epidemia de doença pelo vírus Zika, registrada no país entre 2014 e 2015 que acabou impactando no comportamento reprodutivo das mulheres. Além disso, o país passava por uma acentuada crise econômica e por uma diminuição considerável no número de casamentos. Ainda assim a curva de nascimentos dos anos seguintes sofreu um crescimento considerável.

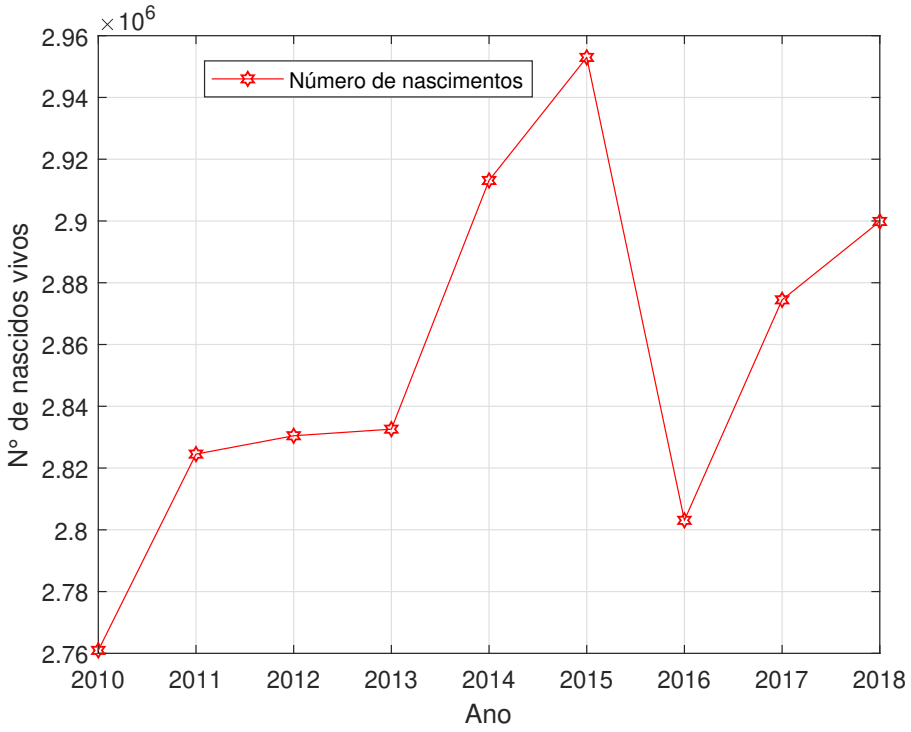



Figura 2.4 – Número de nascidos vivos no Brasil 2010-2018.

### 2.3.2 Superlotação das Unidades

Enquanto muitas maternidades foram sendo fechadas e o número de leitos reduzidos, houve um aumento considerável das taxas de ocupação nas maternidades impulsionadas pelo efeito do crescimento da população e uma maior procura por unidades de saúde ao longo dos últimos anos. Com o aumento das taxas de ocupação, o quadro de superlotação do setor é recorrente, pois a capacidade de atendimento nas maternidades supera a capacidade de acomodação das gestantes nos leitos de internação disponíveis acarretando filas, tempos de espera e permanência elevados, além do aumento das taxas de rejeição de gestantes. O problema de rejeição de gestantes nas redes perinatais não é apenas específico das redes perinatais, mas também é um problema muito comum na maioria das redes de assistência médica, cujo grupo de pacientes alvo requer cuidados urgentes e não podem esperar.

A falta de organização, cooperação e coordenação entre as unidades perinatais pode levar a uma alocação ineficiente dos recursos, políticas improdutivas, indisponibilidade ou inadequação dos recursos necessários, distribuição ineficiente da capacidade que levam à rejeição e desvio de pacientes para outras instalações, mesmo para outras regiões. Esse desvio pode causar longos deslocamentos em busca de outra unidade perinatal, o que pode dar origem a ocorrências fatais. A rejeição pode afetar seriamente o estado de saúde das gestantes e aumentar o risco de mortalidade materna e neonatal. (PEHLIVAN, 2014).

### 2.3.3 Mortalidade Infantil

Diminuir as taxas de mortalidade infantil também é considerado um grande desafio para a saúde pública, pois, apesar da diminuição global de seus índices, ainda é uma realidade presente no Brasil. A Figura 2.5 apresenta as taxas de mortalidade observadas no Brasil durante o período de 2003 a 2018.

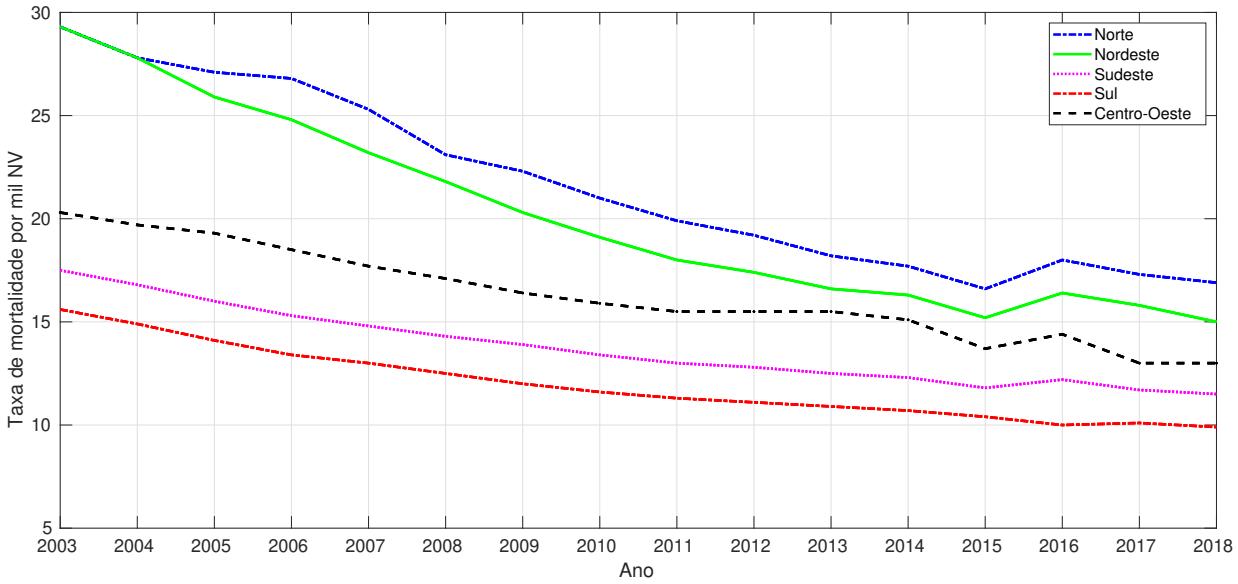



Figura 2.5 – Mortalidade infantil 2003-2018.

Fonte: Sistemas de Informações sobre Mortalidade (SIM) e Sistema de Informações sobre Nascidos Vivos (Sinasc) e Pesquisa de Busca Ativa.

O Boletim Epidemiológico do Brasil, que compreende o período de 2003 a 2019, elaborado pela Secretaria de Vigilância em Saúde (Ministério da Saúde, 2019) mostrou que, entre 2003 e 2018, houve um declínio significativo na taxa de mortalidade infantil no Brasil, de 22,5 em 2003 para 13,1 em 2018 para cada 1.000 nascidos vivos (NV), em todas as regiões do país que pode ser observado na Figura 2.5. No entanto, mesmo com essa redução considerável no número de óbitos infantis nesse período, a proporção de óbitos por causas evitáveis esteve em quase 70%. Os valores médios de mortalidade ainda continuam elevados, sobretudo nas regiões Nordeste e Norte. Ao contrário dos países desenvolvidos, onde predominam as perdas perinatais relacionadas com causas de difícil prevenção, no Brasil as principais causas de óbito perinatal são possíveis de evitar, caso houvesse uma adequada assistência às gestantes durante todo o processo de gravidez até o parto.

### 2.3.4 Localização Geográfica das Unidades de Assistência ao Parto

A localização geográfica de unidades de saúde também é um dos fatores que interferem em sua acessibilidade. O acesso geográfico às maternidades que oferecem infraestrutura de suporte à gestante em trabalho de parto, relacionado à oferta desigual de serviços de saúde de qualidade, é um dos componentes de vulnerabilidade da gestante e do concepto. O cenário de vários municípios periféricos em alguns estados do Brasil contrasta com a superlotação constante enfrentada pelas maternidades das maiores cidades. Alguns munícipios do Brasil não possuem uma infraestrutura mínima de assistência às gestantes, o que acarreta, na maioria dos casos, grandes deslocamentos em busca de um atendimento. Na Figura 2.6 é mostrado o mapa com a distribuição de leitos em cada região do Brasil no mês de junho de 2020 (SAúDE, 2012).

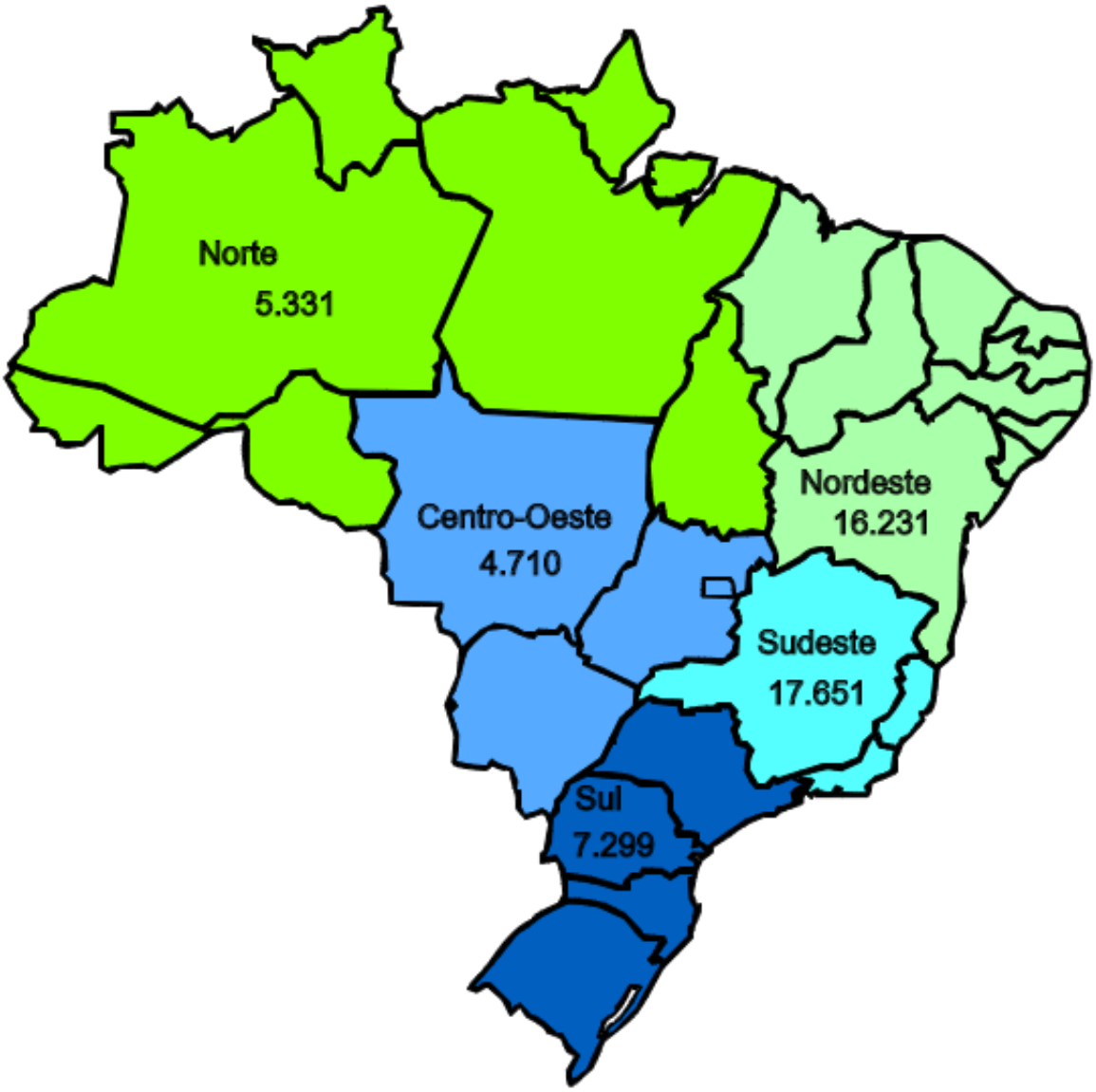



Figura 2.6 – Distribuição de leitos obstétricos por região.

A região norte, a maior região do país, conta com apenas 5331 leitos obstétricos, embora possua a maior taxa de natalidade segundo os dados estimados pelo último censo realizado pelo IBGE. A Tabela 2.3.4 mostra as taxas estimadas de natalidade para cada região do Brasil para o ano de 2019.

Tabela 2.3 – Taxa de natalidade estimada.

| **Região** | **Taxa de Natalidade a cada 1000 habitantes** |
|---|---|
| Norte | 17,06 |
| Nordeste | 14,74 |
| Sudeste | 13,07 |
| Sul | 13,3 |
| Centro-Oeste | 15,3 |

O planejamento da localização de instalações das unidades perinatais é uma estratégia importante que pode melhorar a qualidade e eficiência do atendimento dessas unidades facilitando a prestação de cuidados de saúde. Essa situação dos serviços de obstetrícia é motivo de preocupação, uma vez que observa persistentes taxas de mortalidade perinatal, superlotação nas maternidades, pressão por novos atendimentos e peregrinação das gestantes para encontrar vaga no momento do parto.

O Brasil ainda não possui um sistema de assistência perinatal efetivamente consolidado e regionalizado, embora os indicadores tenham melhorado na última década. O país tem buscado organizar e melhorar a qualidade da assistência perinatal implantando diversas políticas públicas com base em indicadores de saúde sobre a necessidade de cada região do país. No entanto, os serviços de assistência às gestantes e ao parto ainda apresentam baixa qualidade, geralmente por causa de um acompanhamento inadequado às gestantes, à falta de protocolos assistenciais, falta de medicação, de equipamentos para diagnósticos, de equipes especializadas e recursos físicos.

Um sistema regionalizado de cuidados perinatais com parto integrado deve abordar os cuidados recebidos pela mãe antes e durante a gravidez, a gestão do trabalho de parto, os cuidados pós-parto e os cuidados neonatais. Um sistema de saúde que responda às necessidades das famílias, e especialmente das mulheres, requer estratégias para (PRACTICE et al., 2012):

- garantir o acesso aos serviços;
- identificar riscos antecipadamente;
- proporcionar ligação com o nível de cuidado apropriado;
- garantir a adesão, continuidade e abrangência dos cuidados;
- promover o uso eficiente dos recursos.

A organização regionalizada e a integração da assistência perinatal devem evoluir dentro da estrutura do sistema geral de prestação de cuidados de saúde, evitando ao mesmo tempo a duplicação desnecessária de serviços. A integração de um espectro de atividades clínicas, cuidados básicos através de cuidados subespecializados, dentro de um sistema ou região geográfica potencialmente proporciona acesso aos cuidados no nível apropriado para toda a população. O principal objetivo de fornecer o nível adequado de cuidado é facilitado pela avaliação de risco precoce e contínua para prevenir, reconhecer e tratar as condições associadas com morbidade e mortalidade maternas e neonatais. Um objetivo secundário é melhorar encaminhamento e consulta entre instituições que proporcionam diferentes níveis de cuidados. Quando as populações que necessitam de cuidados de saúde reprodutiva são amplamente dispersos, tanto geográfica como economicamente, torna-se necessário um sistema cuidadosamente estruturado e bem organizado de serviços de apoio para garantir o acesso a cuidados apropriados para todas as mulheres grávidas e recém-nascidos. Redes e outras formas de sistemas verticalmente integrados dentro de uma região devem ser estruturados para fornecer todos os serviços necessários, incluindo assistência médica, transporte público e educação profissional, pesquisa e avaliações de resultados com dados organizados em formato padrão. Todos os componentes são necessários para minimizar a mortalidade e morbidade perinatal enquanto os recursos são utilizados de forma eficiente e eficaz (PRACTICE et al., 2012).

### 2.3.5 Taxas Elevadas de Cesáreas

Outro problema enfrentado no Brasil são as elevadas taxas de cesáreas. O ideal, segundo a Organização Mundial de Saúde (OMS) (OMS, 2015), é que o índice de cesáreas se mantenha entre 10 e 15%, bem abaixo do valor encontrado no Brasil. Segundo a OMS, o Brasil detém a segunda maior taxa de cesáreas do mundo com 55%, perdendo apenas para a República Dominicana, onde a taxa é de 56%. Nas unidades da rede pública ainda predomina o parto normal, no entanto, 40% dos partos ocorrem por meio de cirurgias, índice bem superior do que os até 15% recomendados pela OMS. Nas unidades da rede privada, a taxa de cesáreas chega a 84%.

As cesáreas vêm se tornando cada vez mais frequentes tanto nos países desenvolvidos como naqueles em desenvolvimento, principalmente em unidades particulares de saúde (OMS, 2015). Quando realizadas por motivos médicos, as cesáreas podem efetivamente reduzir a mortalidade e a morbidade materna e perinatal (VOGEL et al., 2015). Porém, para casos em que não há complicações ou emergências no momento do parto, as evidências de que fazer cesáreas traga benefícios é contestável. Os partos cirúrgicos são recomendáveis apenas em casos em que o parto natural não pode ser realizado. O parto normal apresenta mais benefícios que as cesáreas, entre eles, uma perspectiva de vida mais saudável para o bebê e redução da mortalidade materna. Especialistas em saúde

consideram o procedimento mais perigoso para as gestantes, com alta taxa de mortalidade, além do tempo de recuperação ser maior. Por ser uma operação, os riscos de hemorragia ou infecção na cesárea são mais comuns quando comparada aos partos normais. No caso dos recém-nascidos, há a possibilidade de se originar problemas respiratórios, diabetes e aumento da pressão sanguínea.

# 3 Algoritmos de Balanceamento de Carga

Modelos de balanceamento de carga estão diretamente ligados à administração de recursos em sistemas distribuídos baseados em modelos de enfileiramento de tarefas e serviços. Os sistemas de enfileiramento surgem naturalmente em muitas aplicações, incluindo redes de comunicação, comutação de pacotes, tráfego de veículos e pessoas, cadeias de produção, entre outros (CRUISE et al., 2020), (SELEN et al., 2016). Essas aplicações são caracterizadas por um intenso fluxo e necessitam de meios para que os sistemas a qual estão inseridos possuam escalabilidade, alta disponibilidade e previsibilidade. Para tornar tais aplicações mais eficientes é necessário projetar políticas de balanceamento de carga que proporcione um bom desempenho diminuindo o tempo de atraso. Isso pode ser feito desenvolvendo ferramentas analíticas para avaliar o desempenho do sistema sob diferentes políticas de balanceamento de carga (GROSU; CHRONOPOULOS, 2005), (CHOW; KOHLER, 1977). Grande parte dos trabalhos desenvolvidos nessa área utilizam tais modelos visando uma alta performance de sistemas que possuem altas taxas de serviços e exigem uma eficiência considerável e operante.

Um sistema distribuído pode ser classificado como homogêneo ou heterogêneo, dependendo das características dos nós constituintes. Um sistema com servidores heterogêneos possui uma taxa de serviço diferente para cada servidor, enquanto que, em um sistema com servidores homogêneos a taxa de serviço é idêntica para todos os servidores. Os sistemas heterogêneos são mais aplicáveis na prática, no entanto, são difíceis de estudar a partir de um ponto de vista matemático. Neste capítulo são apresentados os conceitos básicos sobre os algoritmos de balanceamento de carga, bem como os algoritmos utilizados neste trabalho. É apresentada também a revisão da literatura e o algoritmo de balanceamento proposto.

## 3.1 Trabalhos Relacionados a Balanceamento de Carga

Diversos pesquisadores têm desenvolvido técnicas e estratégias diferentes para lidar com problemas de alocação de recursos e balanceamento de carga em sistemas distribuídos. Alguns destes se preocuparam com o balanceamento de carga em sistemas com servidores homogêneos. Chow e Kohler (CHOW; KOHLER, 1977) compararam o desempenho dos sistemas paralelos *First-Come-First-Served* (FCFS) com dois servidores, em que o tempo entre as chegadas é descrito por uma função exponencial, e os tempos de serviço são analisados sob diferentes políticas de roteamento. Lin e Raghavendra (LIN; RAGHAVENDRA, 1992) propuseram uma política dinâmica de balanceamento de carga com um despachante central de tarefas chamado de *Load-Balancing policy with a Central*

*Job Dispatcher* (LBC) para sistemas distribuídos. Em (LIN; RAGHAVENDRA, 1993) um método de agregação de estados é proposto para analisar o desempenho de políticas dinâmicas de balanceamento de carga e fornece estimativas precisas de desempenho para a política de simetria de sistemas de vários tamanhos quando o atraso médio na transferência de trabalho é pequeno em comparação com o tempo médio de serviço do trabalho.

Eager, Lazowska e Zahorjam (EAGER; LAZOWSKA; ZAHORJAN, 1986) abordaram a questão mais fundamental do nível de complexidade apropriado para políticas de compartilhamento de carga. É mostrado que políticas extremamente simples de compartilhamento de carga adaptativas, que coletam quantidades muito pequenas de informações sobre o estado do sistema e que utilizam essas informações de maneiras muito simples, produzem melhorias dramáticas de desempenho. Em (MIRCHANDANEY; TOWSLEY; STANKOVIC, 1989) estudaram as características de desempenho de algoritmos simples de compartilhamento de carga para sistemas distribuídos e os efeitos de atrasos não negligenciáveis na transferência de tarefas de um nó para outro e na coleta de informações de estado remoto. Os autores analisaram os efeitos desses atrasos sobre o desempenho de três algoritmos, chamados de *forward*, *reverse* e simétricos. Livny e Melman (LIVNY; MELMAN, 1982) apresentaram três diferentes algoritmos de balanceamento de carga para sistemas distribuídos com vários processadores idênticos.

Em (CHOW; KOHLER, 1979) é apresentado o problema de balanceamento de carga em sistemas com servidores heterogêneos. São apresentados, analisados e comparados modelos de enfileiramento para um sistema simples de múltiplos processadores heterogêneos. Cada modelo é diferenciado por uma estratégia de roteamento de tarefas, projetada para reduzir o tempo médio de espera no sistema, equilibrando a carga total entre os processadores. Em cada caso, um trabalho que chega é roteado por um despachante para um dos $m$ processadores paralelos. As estratégias de roteamento de tarefas são divididas em duas classes: determinística e não determinística. As políticas não determinísticas são descritas por probabilidades de ramificação independentes do estado. Para as políticas determinísticas, o próximo processador é escolhido para minimizar ou maximizar o valor esperado de uma função de critério relacionada ao desempenho.

Ni e Hang (NI; HWANG, 1985) consideraram o balanceamento probabilístico de carga em um sistema heterogêneo de múltiplos processadores com muitas classes de trabalho. O esquema de balanceamento de carga é formulado como um problema de programação não linear com restrições lineares. Um algoritmo de balanceamento de carga probabilístico ótimo é proposto para resolver esse problema de programação não linear. Esse método minimiza o tempo médio geral mínimo de resposta do trabalho.

Alguns modelos analisam as políticas de limite com dois servidores, um rápido e outro lento, que alocam um cliente da fila para o servidor rápido sempre que ele estiver disponível, mas usa o servidor lento apenas quando o comprimento da fila excede um

determinado limite. Larsen e Agrawala (LARSEN; AGRAWALA, 1983) estudaram essas políticas para filas infinitas e finitas e mostrou que um nível ótimo de limiar é determinístico e não aleatório. Lin e Kumar (LIN; KUMAR, 1984) mostraram que em sistemas com filas infinitas, a política ideal que minimiza o tempo médio de permanência dos clientes no sistema é do tipo limiar. Stockbridge (STOCKBRIDGE, 1991) analisou sistemas de filas finitas e mostrou que aqui a política de controle ideal nem sempre é do tipo limiar, às vezes é melhor não usar o servidor lento e perder clientes com o bloqueio. Cabral (CABRAL, 2005) estendeu o problema do servidor lento de 2 para $n$ servidores heterogêneos. Para o caso de clientes desinformados ele mostrou que existe um valor da taxa de chegada abaixo do qual o servidor mais lento não deve ser usado e acima do qual deve ser usado.

A política de balanceamento de carga em que as tarefas ingressam na fila mais curta também é estudada e aplicada aos sistemas distribuídos. Nelson e Philips (NELSON; PHILIPS, 1989) derivaram uma aproximação para o tempo médio de resposta de um sistema de múltiplas filas assumindo que existem filas idênticas com capacidade infinita e tempos de serviço que são distribuídos exponencialmente sendo roteadas para uma fila de tamanho mínimo. Lin e Raghavendra (LIN; RAGHAVENDRA, 1996) apresentaram um modelo analítico preciso para avaliar o desempenho da política *Join the Shortest Queue (JSQ)*. O sistema considerado consiste em N filas idênticas, cada uma das quais pode ter um ou mais servidores. Um processo de Markov do tipo nascimento e morte é usado para modelar a evolução do número de trabalhos no sistema. Os resultados obtidos mostraram que este método fornece estimativas muito precisas dos tempos médios de resposta dos postos de trabalho. Gupta et al. (GUPTA et al., 2007) mostraram que para um sistema de compartilhamento de processador com servidores idênticos a política JSQ é quase ideal em termos de minimizar o tempo médio de permanência das tarefas.

Alguns modelos incorporam conceitos dos algoritmos evolutivos. Zomaya e Teh (ZOMAYA; TEH, 2001) investigaram como um algoritmo genético pode ser empregado para resolver o problema do balanceamento dinâmico de carga. Os autores desenvolveram um algoritmo de balanceamento dinâmico que consideram questões de balanceamento de carga, como políticas de limite, critérios de troca de informações e comunicação interprocessadores. Effatparvar e Garshabi (EFFATPARVAR; GARSHASBI, 2014) introduziram um método baseado em algoritmos genéticos para programação e balanceamento de carga em sistemas com multiprocessadores paralelos heterogêneos. Os resultados das simulações indicam que o algoritmo genético reduz o tempo total de resposta e aumenta a utilização.

Grosu, Chronopoulos e Leung (GROSU; CHRONOPOULOS; LEUNG, 2002) formularam o problema de balanceamento de carga em sistemas de trabalho de classe única distribuídos como um jogo cooperativo entre computadores. Subrata, Zomaya e Landfeldt (SUBRATA; ZOMAYA; LANDFELDT, 2007) propuseram uma solução teórica de jogo para o problema de balanceamento de carga da rede. Os autores modelaram o problema

de balanceamento de carga da rede como um jogo não cooperativo, no qual o objetivo é alcançar o equilíbrio de *Nash*.

Diante dos trabalhos apresentados nota-se que, existem diversas políticas de balanceamento de carga que foram propostas ao longo dos anos e que se baseiam em diferentes técnicas de controle e roteamento de tarefas ou clientes visando maior eficiência nos sistemas distribuídos. Embora os diversos modelos e algoritmos de balanceamento de carga apresentados estejam diretamente ligados à administração de recursos em sistemas distribuídos computacionais, tais modelos podem ser aplicados a outros sistemas e redes que utilizem enfileiramentos, como é o caso dos sistemas de saúde. Os centros de saúde em geral possuem uma organização sistêmica que as configuram como um sistema distribuído. Os hospitais possuem diferentes unidades internas, com diversas especialidades médicas interagindo entre si. Assim como nos sistemas distribuídos computacionais, nos sistemas de saúde um dos principais objetivos a ser alcançado é encontrar a melhor política operacional para minimizar o tempo médio de permanência dos clientes ou pacientes no sistema. Como a demanda por serviços de saúde cresce à medida que a população aumenta, é necessário procurar ferramentas operacionais que contribuam para uma distribuição justa e eficiente dos recursos de saúde.

## 3.2 Balanceamento de Carga Inspirado no Comportamento das Abelhas

Nesta seção é apresentado um problema de programação linear inteira mista onde é utilizado um algoritmo de balanceamento de carga baseado na meta-heurística do comportamento das abelhas melíferas.

### 3.2.1 Inteligência de Enxame

A inteligência de enxames, também referenciada como inteligência de colônias ou inteligência coletiva, é um conjunto de técnicas baseadas no comportamento coletivo de sistemas auto-organizados, distribuídos, autônomos, flexíveis e dinâmicos. Estes sistemas são formados por uma população de agentes que têm a capacidade de perceber e modificar seu ambiente local (SERAPIÃO, 2009).

Bonabeau et al. (BONABEAU et al., 1999) definiram a inteligência de enxame como qualquer tentativa de projeto de algoritmos ou dispositivos distribuídos para solução de problemas inspirados no comportamento coletivo de colônias sociais de insetos e outras sociedades animais. Eles focalizaram seu ponto de vista somente em insetos sociais, como cupins, abelhas, vespas e outras espécies diferentes de formigas. No entanto, o termo enxame é usado de forma geral para se referir a qualquer coleção restrita de agentes ou indivíduos

que interagem (KARABOGA, 2005). Tem-se, na organização de abelhas em torno de sua colmeia, um exemplo clássico de enxame. Demais sistemas que possuem uma arquitetura similar podem ser compreendidos como enxames por possuírem o comportamento coletivo auto-organizado. Outros exemplos de enxames podem ser observados nas colônias de formigas cujos agentes individuais são formigas e em um bando de aves que pode ser considerado um enxame de pássaros.

A otimização por colônia de formigas, *Ant Colony Optimization* (ACO), trabalha com sistemas artificiais inspirados no comportamento forrageiro de formigas reais, que são usados para resolver problemas discretos de otimização (DORIGO; MANIEZZO; COLORNI, 1996). A ideia principal é a comunicação indireta entre as formigas que liberam feromônio durante seu trajeto em busca de comida, o que lhes permite encontrar caminhos curtos entre seu ninho e a comida. Existem diversos algoritmos diferentes que podem ser considerados como otimização por colônia de formigas (Gupta; Deshpande, 2014), (LI et al., 2011), (NISHANT et al., 2012), (KESKINTURK; YILDIRIM; BARUT, 2012), (GAMBARDELLA; TAILLARD; AGAZZI, 1999), mas o primeiro deles, chamado *Ant System*, foi proposto por (DORIGO, 1992). Inicialmente três diferentes modelos foram propostos: *AS-density*, *AS-quantity* e *AS-cycle*, diferindo-se pela maneira que as trilhas de feromônio eram atualizadas (DORIGO; MANIEZZO; COLORNI, 1996). O último modelo mostrou-se mais eficiente e a maior parte dos algoritmos de otimização por colônia de formigas atuais derivam-se dele.

Têm-se também algoritmos que utilizam a otimização de enxame de partículas, *Particle Swarm Optimization* (PSO), que incorporam e modelam os comportamentos de enxameação observados em bandos de aves, cardumes de peixes e até mesmo o comportamento social humano (PARSOPOULOS; VRAHATIS, 2004), (VESTERSTRØM; RIGET, 2002), (CLERC; KENNEDY, 2002). A otimização PSO é uma ferramenta de otimização baseada em população, que pode ser implementada para resolver problemas de otimização que podem ser definidos ou aproximados por meio de funções (ABRAHAM; GUO; LIU, 2006).

Um sistema imunológico também é considerado um tipo de enxame onde seus agentes individuais são compostos por células e moléculas. Esse conceito de enxame de células e moléculas têm sido utilizado pela engenharia imunológica com o objetivo de criar ferramentas para resolver problemas de aprendizagem de máquina, utilizando informações extraídas dos próprios problemas (CASTRO; ZUBEN, 1999), (SMITH et al., 1998), (HIGHTOWER; FORREST; PERELSON, 1995).

Um dos principais elementos da inteligência de enxame é a auto-organização. Segundo (BONABEAU et al., 1999) a auto-organização em enxames pode ser interpretada através de quatro características:

1. *Feedback* positivo: promovendo a criação de estruturas convenientes. Um exemplo de feedback positivo se dá no recrutamento e reforço como a colocação de trilhas e acompanhamento em algumas espécies de formigas ou danças em abelhas;

2. *Feedback* negativo: contrabalançando o feedback positivo e ajudando a estabilizar o padrão coletivo. Pode tomar a forma de saturação, exaustão ou competição. No exemplo de forrageamento, o feedback negativo deriva do número limitado de forrageadoras disponíveis, da saciedade, da exaustão da fonte de alimento, da aglomeração na fonte de alimento ou competição entre fontes de alimento;

3. Flutuações: caminhadas aleatórias, erros, troca de tarefas aleatórias entre os indivíduos do enxame, que são vitais para a criatividade. A aleatoriedade é frequentemente significativa para estruturas emergentes, uma vez que permite a descoberta de novas soluções;

4. Interações múltiplas: os agentes do enxame utilizam as informações provenientes de outros agentes para que as informações se espalhem por toda a rede.

Além dessas características, um enxame inteligente possui a capacidade de executar tarefas simultaneamente por agentes especializados, dividindo o trabalho entre seus agentes. De acordo com (MILLONAS, 1994), as principais propriedades de um sistema de inteligência de enxame são:

- Proximidade: garante que os agentes possam interagir fazendo cálculos simples de espaço e tempo;

- Qualidade: garante que os agentes possam avaliar seu comportamento respondendo a fatores de qualidade no meio ambiente;

- Diversidade: permite que o sistema reaja a situações inesperadas;

- Estabilidade: o enxame não deve mudar seu modo de comportamento a cada variação do ambiente;

- Adaptabilidade: o enxame deve ser capaz de mudar seu modo de comportamento quando necessário.

A inteligência dos enxames também é considerada um ramo da abordagem computacional conhecida como Computação Natural. Esta abordagem é baseada no processo de extração de ideias da natureza para desenvolver sistemas computacionais (SERAPIÃO, 2009). A Computação Natural é indicada em técnicas de otimização quando se tem problemas complexos, com um grande número de variáveis e diversas soluções possíveis, é altamente dinâmica, não linear e com múltiplos objetivos. Geralmente esses problemas

não garantem que a solução encontrada seja ótima, mas é possível encontrar uma medida de qualidade que permita a comparação entre soluções diversas (CASTRO, 2006).

### 3.2.2 Algoritmos de Colônias de Abelhas

Uma colônia de abelhas pode ser classificada como uma sociedade de inteligência coletiva, onde várias atividades são realizadas em conjunto para atingir um objetivo específico que beneficia toda a população. Esta coleção de agentes, formada por abelhas, pode ser definida como um enxame inteligente, pois possui vários agentes ou indivíduos interagindo entre si e características como auto-organização e divisão de trabalho estão presentes (KARABOGA, 2005). Esse comportamento inteligente das colônias de abelhas com suas características de aprendizagem, memorização e compartilhamento de informações têm sido alvo de diversas pesquisas sobre inteligência de enxames. Muitos pesquisadores desenvolveram nas últimas décadas, algoritmos baseados no comportamento das abelhas que proporcionaram diversas abordagens diferentes e algoritmos distintos sob as características comportamentais dessas colônias.

Inspirados pelo comportamento inteligente de enxames de abelhas em busca de alimento ou forrageamento, Lucic e Teodorovic (LUCIC; TEODOROVIć, 2001) introduziram o algoritmo *Bee System* (BS) na solução de problemas de otimização combinatorial. O algoritmo BS foi usado para resolver o problema do caixeiro viajante. Abbass (Abbass, 2001) desenvolveu um algoritmo de otimização baseado no processo de acasalamento das abelhas. Nakrani e Tovey (NAKRANI; TOVEY, 2003) desenvolveram o algoritmo *Honey Bee Algorithm* (HBA) para realizar alocação de servidores na Internet.

Sung (JUNG, 2003) propôs um método de evolução abelha-rainha chamado *Queen-Bee Evolution for Genetic Algorithms* (QEGA) com o objetivo de melhorar a capacidade de otimização dos algoritmos genéticos. Wedde, Farooq e Zhang (WEDDE; FAROOQ; ZHANG, 2004) apresentaram um algoritmo de roteamento chamado *BeeHive*, que foi inspirado na comunicação, métodos e procedimentos de avaliação de abelhas melíferas.

Bozorg Haddad e Ashfar (HADDAD; AFSHAR, 2004) desenvolveram um algoritmo baseado no vôo nupcial das abelhas, conhecido como *Mating Bee Optimization* (MBO) aplicado ao problema de otimização de reservatório usando variáveis discretas, que posteriormente foi melhorado para o algoritmo *Honey-Bee Mating Optimization* (HBMO) utilizado em otimização de funções contínuas (HADDAD; AFSHAR; MARINO, 2006). Yang (YANG, 2005) desenvolveu o algoritmo *Virtual Bee Algorithm* (VBA) para otimização de funções contínuas de duas variáveis em problemas de engenharia.

Karaboga (KARABOGA, 2005) propôs o algoritmo *Artificial Bee Colony Algorithm* (ABC), uma meta-heurística baseada na população de uma colônia de abelhas artificiais para a otimização multidimensional e multimodal de problemas. Pham et al. (PHAM et

al., 2006) desenvolveram um algoritmo de otimização inspirado no comportamento de coleta de alimentos de abelhas produtoras de mel, chamado *Bees Algorithm* (BA) que pode ser usado tanto para otimização combinatória quanto para otimização funcional.

Lu e Zhou (LU; ZHOU, 2008) desenvolveram o algoritmo *Bee Collecting Pollen Algorithm* (BCPA) inspirados no comportamento das abelhas coletoras de pólen para problemas de convergência global. Posteriormente, Bernardino et al. (BERNARDINO et al., 2011) propuseram um balanceamento de carga baseado no comportamento das abelhas, aplicando-o ao *Resilient Packet Ring* (RPR), também conhecido como IEEE 802.17, que é um padrão projetado para otimizar o tráfego de dados em redes de fibra ótica com topologia em anel. Babu e Krishna (BABU; KRISHNA, 2013) também desenvolveram, inspirados no comportamento das abelhas, o algoritmo *Honey bee Behavior Inspired Load Balancing* (HBB-LB) para o balanceamento de carga de tarefas em um ambiente de computação em nuvem.

### 3.2.3 Comportamento do Enxame de Abelhas

As abelhas melíferas são insetos sociais, o que significa que vivem em conjunto em grandes grupos familiares bem organizados. Uma colônia de abelhas consiste tipicamente de três tipos de abelhas adultas: operárias, zangões e uma rainha. Vários milhares de abelhas operárias cooperam na comunicação, na construção de ninhos complexos, na coleta de alimentos, na defesa da colmeia, no controle ambiental e na divisão do trabalho que permite às abelhas sucesso nas colônias sociais. Para sobreviverem, as colônias dependem dessa diversidade da população, já que cada classe de abelhas realiza tarefas específicas funcionando como um sistema dinâmico que se ajusta de acordo com suas necessidades, ainda que seus agentes, as abelhas, se observados de forma individual possuam capacidade e conhecimento limitados. Assim, embora as rainhas sejam extremamente poderosas dentro de suas sociedades, elas não podem estabelecer novas colônias sem a ajuda de zangões e operárias, que fornecem fertilização, alimento e material para a construção da colmeia.

Cada colônia tem apenas uma rainha, exceto durante um período variável após os preparativos para a enxameação ou a substituição. Sua função primária é a reprodução. A segunda função principal de uma rainha é produzir feromonas que servem como uma identificação social unificando e ajudando a dar identidade individual a uma colônia de abelhas. Os zângãos (abelhas machos) são as maiores abelhas da colônia. Elas estão geralmente presentes apenas no final da primavera e no verão. Sua principal função é fertilizar a rainha virgem durante seu voo de acasalamento. As operárias são as menores e constituem a maioria das abelhas que ocupam a colônia. São fêmeas sexualmente não desenvolvidas e, em condições normais de colmeia, não põem ovos. As operárias possuem estruturas especializadas, tais como glândulas de criação, glândulas de cheiro, glândulas de cera e cestas de pólen, que lhes permitem realizar todo o trabalho da colmeia.

Uma das atividades mais importantes para a subsistência da colmeia é o forrageamento. O forrageamento consiste na busca por novas fontes de alimento ou exploração das fontes já conhecidas que exigem o trabalho em conjunto da colmeia, seja para intensificar a exploração de boas fontes ou para o abandono de fontes com recursos esgotados. Nas colônias, a classe das abelhas operárias (campeiras) é responsável pela busca de alimentos. Quando encontram uma fonte de alimento, estas abelhas retornam à colmeia para anunciar a descoberta utilizando uma dança peculiar. Por meio dessa dança as operárias podem informar a qualidade e a quantidade dos alimentos além da distância e localização exata dessa fonte de alimento. Estas informações ajudam a colônia a enviar suas abelhas para fontes de alimento com certa precisão sem utilizar guias ou mapas. Depois de anunciar a nova fonte de alimento as abelhas operárias seguidoras seguem as abelhas operárias campeiras até a fonte de alimento e começam a colhê-la.

O tipo de dança que a abelha realiza depende da distância da colmeia à fonte de alimento que é diferenciada pelo número de vibrações realizadas e pela intensidade do som emitido durante a dança. Quanto menor a distância entre a fonte e a colmeia, maior o número de vibrações. Portanto, as danças indicam que há uma fonte abundante de alimento, os movimentos indicam a distância e a orientação, e o néctar ou pólen passado para as recrutas ajudam a reconhecer a fonte pelo odor do alimento (SERAPIÃO, 2009).

Cada abelha pode se comportar de três maneiras diferentes depois de descarregar os alimentos: ela pode fazer a dança das abelhas para recrutar mais abelhas forrageiras para a mesma fonte de alimentos; ela pode abandonar a fonte de alimentos devido à falta de recursos disponíveis; ou pode voltar diretamente à forragem. A ideia básica relativa aos algoritmos baseados no comportamento das abelhas forrageiras é que as abelhas forrageiras têm uma potencial solução para um problema de otimização em sua memória (ou seja, uma configuração para as variáveis de decisão do problema). Esta solução potencial corresponde à localização de uma fonte de alimento e tem uma medida de qualidade agregada (ou seja, valor da função objetivo). As informações de qualidade da fonte alimentar são trocadas através da dança que probabilisticamente viesa outras abelhas para explorar fontes de alimentos com maior qualidade (PARPINELLI; LOPES, 2011).

### 3.2.4 O Algoritmo de Balanceamento de Carga Aplicado ao Sistema Perinatal

A admissão de uma gestante em uma maternidade pode ser comparada a uma abelha que procura uma fonte de alimento (uma flor ou um canteiro de flores). O alimento na fonte pode ser escasso, e uma nova fonte de alimento pode ser necessária. Na analogia proposta, o serviço de cuidados para gestantes que está sendo feito é semelhante ao processo de esgotamento de alimentos em uma única fonte. Quando uma maternidade está sobrecarregada, a gestante será transferida para uma maternidade com capacidade de atendimento disponível, como uma abelha operária campeira que encontra uma nova fonte

de alimento. Esta transferência é responsável pela atualização do sistema composto por gestantes que estão esperando na fila de cuidados, utilizando uma etapa análoga à dança realizada pelas abelhas operárias campeiras ao informar às abelhas operárias seguidoras que encontraram alimentos. Esta atualização proporcionará subsídios para decidir em qual maternidade a gestante deve ser internada com o objetivo de obter uma carga bem balanceada em todas as maternidades.

#### 3.2.4.1 Modelo Matemático

O modelo matemático para o analogia proposta é apresentado da seguinte maneira: seja M = $\{M_1, M_2, ..., M_j..., M_m\}$ o conjunto de $m$ maternidades, $\mu = \{\mu_1, \mu_2, ..., \mu_j, ..., \mu_m\ \}$, o conjunto com as suas $m$ respectivas taxas de atendimento, que deverão atender $n$ gestantes representado por P = $\{P_1, P_2, ..., P_i, ...P_n\}$. O tempo de espera $TT_{ij}$ é definido como a soma do tempo de espera $WT_{ij}$ na fila e o tempo de atendimento $CT_{ij}$ de uma gestante $P_i$ na maternidade $M_j$, como segue:

$$TT_{ij} = \sum_{j=1}^{m}\sum_{i=1}^{n}(CT_{ij}.a_{ij} + WT_{ij}.b_{ij}) \tag{3.1}$$

As seguintes variáveis de decisão binárias inteiras são definidas:

$$a_{ij} = \begin{cases} 1, \text{ se a gestante } P_i \text{ está admitida na maternidade } M_j \\ 0, \text{ caso contrário} \end{cases}$$

$$b_{ij} = \begin{cases} 1, \text{ se a gestante } P_i \text{ está na fila de atendimento da maternidade } M_j \\ 0, \text{ caso contrário} \end{cases}$$

O tempo médio de permanência $TT$ é definido em (3.2) pela soma dos $TT_{ij}$ divididos pela quantidade total de gestantes $QP$ que já foram atendidas em todas as maternidades do sistema perinatal:

$$TT = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n_j} TT_{ij}}{QP}. \tag{3.2}$$

onde $n_j$ é a quantidade total de gestantes na maternidade $j$ e $QP$ é definido como segue:

$$QP = \sum_{j=1}^{m} n_j. \tag{3.3}$$

Portanto, o objetivo é reduzir $TT$ no sistema perinatal representado pela função objetivo dada em (3.5). A formulação do problema é a seguinte:

$$\text{Minimize} \quad TT \tag{3.4}$$

$$\text{Sujeito a:} \quad WT_{ij} = \sum_{z=1}^{WQ}(min(CT_{jz})) \tag{3.5}$$

$$1 < CT_{ij} \leq 24 \quad i = 1, 2, ..., n, \quad j = 1, 2, ..., m \tag{3.6}$$

$$a_{ij} \in \{0, 1\}, b_{ij} \in \{0, 1\} \quad i = 1, 2, ..., n, \quad j = 1, 2, ..., m \tag{3.7}$$

Em (3.5), o tempo de espera ($WT$) é calculado em função do tempo de atendimento ($CT$) do número de gestantes na fila de espera ($WQ$). O $WT_{ij}$ de cada $P_i$ em $M_j$ é calculado levando em conta o menor tempo de atendimento ($CT$) das $M_j$ para cada gestante. Em (3.6), o $CT_{ij}$ que é a taxa de atendimento média pode ser descrita por uma distribuição exponencial com o parâmetro $\mu_j$, que é totalmente determinada de acordo com a disponibilidade de pessoal na maternidade e sua quantidade de leitos.

O $CT$ é um dos principais fatores para calcular o tempo de espera em uma fila em uma abordagem de tempo discreto. Em (3.7), as variáveis de decisão inteira são binárias e usadas para indicar a admissão da gestante na maternidade. Para uma melhor compreensão, resumimos os parâmetros do sistema na Tabela 3.1.

Tabela 3.1 – Parâmetros do Sistema.

| **Símbolo** | **Descrição** |
|---|---|
| $\mu_j$ | Taxa de atendimento da maternidade |
| $\lambda_j$ | Taxa de chegada de gestantes |
| $TT$ | Tempo médio de permanência |
| $P_i$ | Gestante |
| $M_j$ | Maternidade |
| $WT_{ij}$ | Tempo de espera |
| $CT_{ij}$ | Tempo de atendimento |
| $WQ$ | Gestantes na fila de espera |
| $CP$ | Gestantes atendidas |
| $QP$ | Total de gestantes atendidas |
| $M_o$ | Maternidades sobrecarregadas |
| $M_u$ | Maternidades com baixa carga |
| $SB$ | Sistema Balanceado |

O problema é NP-difícil, pois apresenta uma natureza combinatória que dificulta encontrar a solução ótima com eficiência computacional. O esforço computacional envolvido neste tipo de problema reside nas variáveis binárias inteiras, o que torna os métodos meta-heurísticos atraentes para encontrar boas soluções (BANDYOPADHYAY, 2007). A taxa de chegada total de gestantes no sistema perinatal é definida por uma distribuição de Poisson com o parâmetro $\lambda$ (BRANDEAU; SAINFORT; PIERSKALLA, 2004), onde $\lambda$ no sistema é definido como a soma das taxas de chegada de gestantes nas maternidades, ou seja, $\lambda = \lambda_1 + \lambda_2 + ... + \lambda_m$.

#### 3.2.4.2 O Algoritmo

A carga é definida pelo número de gestantes em atendimento ou espera na fila da maternidade do hospital $M_j$. Todas as maternidades serão classificadas em ordem ascendente e descendente e agrupadas com base em sua carga, conforme descrito pelo **Algoritmo 1**. As maternidades sobrecarregadas $M_o$ são denotadas pelo número de gestantes maior que a capacidade de atendimento. Maternidades com baixa carga $M_u$ são denotadas pelo número de gestantes menores que a capacidade de atendimento.

**Algoritmo 1:** Maternidades agrupadas com base em sua carga

```
1  p = Poisson(λ);
2  e = Exp(μ);
3  para j de 1 até length(M) faça
4      se p_j > e_j então
5          M_o = M_j;
6      fim
7      senão se p_j < e_j então
8          M_u = M_j;
9      fim
10     se SB; então
11     fim
12 fim
```

O sistema deve decidir se deseja fazer o balanceamento de carga ou não. Para isso, há duas situações possíveis:

- primeiro, descobrir se o sistema está equilibrado;
- segundo, descobrir se o sistema está desequilibrado e se tem gestantes na fila.

Se uma gestante não encontrar leitos na maternidade, ela esperará na fila até ser atendida. As gestantes atendidas são denotadas por $CP$. No momento do balanceamento de carga, as gestantes serão transferidas de uma maternidade para outra, a fim de reduzir o tempo médio de permanência no sistema. Quando uma maternidade fica sobrecarregada, o ciclo começa até que todas as gestantes sejam admitidas nas maternidades e o sistema seja balanceado com base em sua capacidade. Esta técnica de balanceamento de carga é

descrita pelo **Algoritmo 2**.

**Algoritmo 2:** Balanceamento de Carga Inspirado no Comportamento das Abelhas

**Entrada :** $\mu_1$, $\mu_2$, $\lambda_1$, $\lambda_2$;
**Inicialize:** $CT = 0$, $WT = 0$, $TT = 0$;
Ordena as $M_o$ em ordem decrescente e as $M_u$ em ordem crescente;
$z_1 = Poisson(\lambda_1)$;
$z_2 = Poisson(\lambda_2)$;
$z = z_1 + z_2$;
**para** *i de 1 até length(z)* **faça**
    $CP = (z_1(i) + z_2(i))$;
    $TP = TP + CP$;
    **enquanto** $WQ$ *or* $M_o$ **faça**
        $WT = WT + min(CT)$;
        **repita**
            **se** $M_o$ *and* $M_u$ **então**
                Remova a gestante da $M_o$;
                Admita a gestante na $M_u$;
            **fim**
            **senão se** $WQ$ *and* $M_u$ **então**
                Remova a gestante da $WQ$;
                Admita a gestante na $M_u$;
            **fim**
            **se** *SB;* **então**
            **fim**
            Ordena as $Mo$ em decrescente e as $Mu$ em crescente;
        **até** $SB$;
        $CT = CT + exprnd(24, [1, AP])$;
        Esvazia as maternidades retirando as $CP$;
        Ordena as $Mo$ em ordem decrescente e as $Mu$ em ordem crescente;
    **fim**
**fim**
**Saida:** $TT = (CT + WT)/QP$;

## 3.3 Balanceamento Dinâmico de Carga

Nesta seção são apresentados diferentes modelos de balanceamento de carga propostos por (CHOW; KOHLER, 1979). Cada modelo é diferenciado por uma estratégia de roteamento de tarefas projetada para reduzir o tempo médio de espera no sistema

*(Turnaround Time)*. Em (CHOW; KOHLER, 1979) são utilizadas duas classes de roteamento: não determinística e determinística. O tempo médio de permanência no sistema é calculado utilizando teoria de filas para a classe não determinística e uma cadeia de Markov de tempo contínuo e estado discreto para a classe determinística.

### 3.3.1 Roteamento Não Determinístico

Considere o modelo de sistema heterogêneo de múltiplos processadores na Figura 3.1. O sistema é composto por $m$ servidores, com tempo de processamento exponencial e taxa média de atendimento $\mu_i$ , assim o processo de saída de tarefas do sistema segue um processo de Poisson. O tempo entre as chegadas de tarefas também é caracterizado por uma distribuição exponencial com taxa média $\lambda$. A disciplina seguida pelas filas é do tipo (*First-Come-First-Served*), ou seja, as tarefas são processadas com base na ordem de chegada, a primeira a chegar a uma fila é a primeira a ser processada pelo servidor dessa fila. O sistema é aberto e, portanto, as tarefas que saem do sistema não voltam a ingressá-lo.

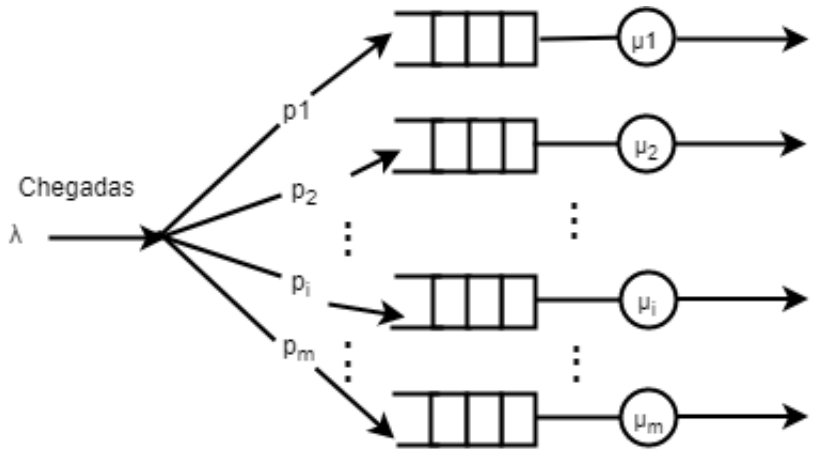



Figura 3.1 – Sistema heterogêneo com roteamento não determinístico (CHOW; KOHLER, 1979).

O comportamento do sistema pode ser descrito por um processo de Markov de tempo contínuo e estado discreto, onde $s(n_1, n_2, ..., n_i, ..., n_m)$ é o estado do sistema e $n_i$ é o número de tarefas da $i$-ésima fila, incluindo a tarefa em serviço. Cada tarefa que chega é roteada para uma das filas com probabilidade $p_i$. Dessa forma, tem-se:

$$\sum_{i=1}^{m} p_i = 1. \tag{3.8}$$

Se a probabilidade de roteamento $p_i$ é fixa, o sistema apresentado na Figura 3.1 pode ser decomposto e analisado como $m$ sistemas M/M/1 independentes, com taxa de

chegada $\lambda p_i$ e taxa de atendimento $\mu_i$, sendo $\lambda p_i < \mu_i$ para que o sistema permaneça estável. O tempo médio de permanência no sistema (TT: *Turnaround Time*) é dado por:

$$\begin{aligned} TT &= \sum_{i=1}^{m} p \left( \frac{1}{\mu_i - \lambda \rho_i} \right) \\ &= \frac{m}{\sum_{j=1}^{m} \mu_j - \lambda}. \end{aligned} \tag{3.9}$$

### 3.3.2 Roteamento Determinístico

Considere o modelo de sistema heterogêneo de múltiplos processadores na Figura 3.2. Este modelo é o mesmo da Figura 3.1, exceto pela inclusão do despachante de tarefas. Quando chega ao sistema, a tarefa é encaminhada, pelo despachante, para a fila $q(s, C)$, onde $s(n_1, n_2, ..., n_m)$ é o estado do sistema, $n_1$, $n_2$, ..., $n_m$ é o número de usuários no sistema e $C$ é a função de critério utilizada pelas estratégias de roteamento.

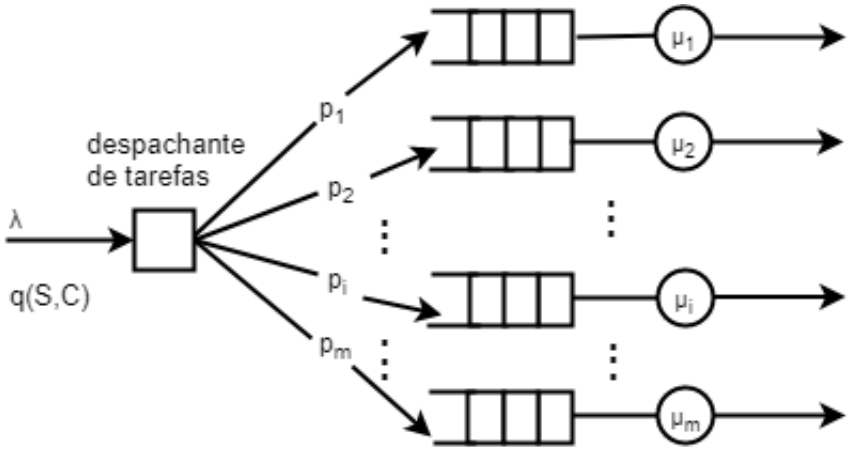



Figura 3.2 – Sistema heterogêneo com roteamento determinístico (CHOW; KOHLER, 1979).

A estratégia de roteamento de tarefas é determinística, uma vez que uma tarefa de chegada é sempre roteada para um processador de acordo com a função de critério determinística do despachante de tarefas. As características do sistema com modelagem determinística são similares aos do sistema com modelagem não determinística, isto é, as distribuições que caracterizam o tempo entre chegadas e o tempo de processamento das tarefas pelos servidores são dados por uma exponencial, a disciplina seguida pelas filas é do tipo FCFS e o sistema é aberto. Serão investigadas três políticas de roteamento dependentes do estado: a política de tempo mínimo de resposta, a política de tempo mínimo do sistema e a política de máxima vazão. Estas políticas foram propostas por (CHOW; KOHLER, 1979).

#### 3.3.2.1 Política de Tempo Mínimo de Resposta

A política de tempo mínimo de resposta direciona uma tarefa que chega ao sistema para a fila que oferece menor tempo médio de permanência. A política pode ser declarada formalmente da seguinte maneira:

1. A tarefa é enviada para a fila que apresente após a chegada da próxima tarefa a menor relação entre o tamanho da fila (incluindo a tarefa no servidor–HOL (*Head-of-Line*)) e a taxa de serviço:

$$\frac{n_i + 1}{\mu_i} = min\left\{\frac{n_k + 1}{\mu_k}, k = 1, 2, ..., m\right\}. \tag{3.10}$$

2. Se a relação não for única, a fila com maior taxa $\mu_i$ é selecionada.

Note que assumiu-se $\mu_i \neq \mu_j$ para $i \neq j$. A política de tempo mínimo de resposta é baseada na visão de um usuário individual de como melhorar o desempenho, uma vez que $\left(\frac{n_k+1}{\mu_k}\right)$ é o tempo de resposta esperado da próxima tarefa no sistema $k$. A política de tempo mínimo de reposta é definida formalmente a seguir.

*Política de Tempo Mínimo de Resposta (R):*

$R(n_1, n_2, ..., n_{k+1}, ..., n_m) \triangleq \frac{n_k+1}{\mu_k}$

$R_{min}(\text{s}) \triangleq \min\{R(n_1, n_2, ..., n_{k+1}, ...n_m | k = 1, 2, ..., m)\}$

$Q(s, R) \triangleq \min\{k | R_{min}(s) = R(n_1, n_2, ..., n_{k+1}, ...n_m | k = 1, 2, ..., m)\}$.

#### 3.3.2.2 Política de Tempo Mínimo do Sistema

A política de tempo mínimo do sistema busca minimizar o tempo de atendimento de todas as tarefas que se encontram no sistema. É motivada para encontrar a estratégia de roteamento ideal e para modelar o processamento paralelo. Dado que o sistema está no estado $s(n_1, n_2, ..., n_i, ..., n_m)$, o tempo do sistema $T(n_1, n_2, ..., n_i, ..., n_m)$ é definido como o tempo esperado para concluir todos os trabalhos já existentes no sistema. Esta definição pode ser expressa recursivamente da seguinte maneira:

$$\begin{aligned} T(n_1,\ n_2,\ ...,\ n_m) &= E[\textit{tempo para atender todas as tarefas}] \\ &= E[\textit{tempo para atender a próxima tarefa}] \\ &\quad + E[\textit{tempo para atender as demais tarefas}] \\ &= \frac{1}{\sum_{\substack{j=1 \\ n_j \neq 0}}^{m} \mu_j} + \sum_{\substack{k=1 \\ n_k \neq 0}}^{m} \left[\frac{\mu_k}{\sum_{\substack{j=1 \\ n_j \neq 0}}^{m} \mu_j} T(n_1, n_2, ..., n_{k-1}, ..., n_m)\right]. \end{aligned} \tag{3.11}$$

O termo $1/\sum_{j=1, n_j \neq 0}^{m} \mu_j$ representa o tempo médio até a próxima tarefa concluir seu serviço, dado que o sistema está no estado $(n_1, n_2, ...n_m)$. Isso decorre da suposição de que

cada servidor tem uma distribuição de serviço exponencial. A expressão $\mu_k / \sum_{j=1,n_j \neq 0}^{m} \mu_j$ é a probabilidade de que a próxima tarefa de saída seja da fila $k$. Se T$(n_1, n_2, ..., n_m)$ for escolhido como a função de critério, a política de tempo mínimo do sistema é definida formalmente da seguinte maneira.

*Política de Tempo Mínimo de Sistema (T):*

$T_{min}(s) \triangleq \min\{T(n_1, n_2, ..., n_{k+1}, ...n_m | k = 1, 2, ..., m)\}$

$Q(s, R) \triangleq \min\{k | T_{min}(s) = R(n_1, n_2, ..., n_{k+1}, ...n_m | k = 1, 2, ..., m)\}$

Então $q(s, R) = i$, onde $i \in Q(s, R)$ e $\mu_i < \mu_j \ \forall j \in Q(s, R), \ j \neq i$.

#### 3.3.2.3 Política de Máxima Vazão

O tempo mínimo de resposta e as políticas de tempo do sistema são apenas funções dos comprimentos de fila do sistema e taxas de serviço do processador. A política de máxima vazão leva em consideração, também, a taxa de chegadas $\lambda$, com o objetivo de melhorar o desempenho do sistema.

Dado que o sistema está no estado $(n_1, n_2, ...n_m)$, a vazão média geral $TP(n_1, n_2, ...n_m)$. durante o próximo período de comparação é a soma de todos os $TP_i(n_i)$ individuais. Isso é dado por:

$$TP(n_1, n_2, ...n_m) = \sum_{i=1}^{m} \lambda \left[ \sum_{k=1}^{n_i - 1} \left(1 - \frac{n_i}{k}\right) \left(\frac{\mu_i}{\lambda + \mu_i}\right)^k - \mu_i ln \left(\frac{\lambda}{\lambda + \mu_i}\right) \right]. \tag{3.12}$$

A política de máxima vazão é definida formalmente a seguir.

*Política de Máxima Vazão (TP):*

$TP_{max}(s) \triangleq \max\{TP(n_1, n_2, ..., n_{k+1}, ...n_m | k = 1, 2, ..., m)\}$

$Q(s, R) \triangleq \min\{k | T_{max}(s) = TP(n_1, n_2, ..., n_{k+1}, ...n_m | k = 1, 2, ..., m)\}$

Então $q(s, R) = i$, onde $i \in Q(s, R)$ e $\mu_i < \mu_j \ \forall j \in Q(s, R), \ j \neq i$.

### 3.3.3 Técnica de Solução Recursiva

Para analisar os métodos determinísticos apresentados, Chow e Kohler (CHOW; KOHLER, 1979) propõem um algoritmo recursivo para cálculo da matriz de probabilidades de estado de equilíbrio (ou estacionário) para dois processadores.

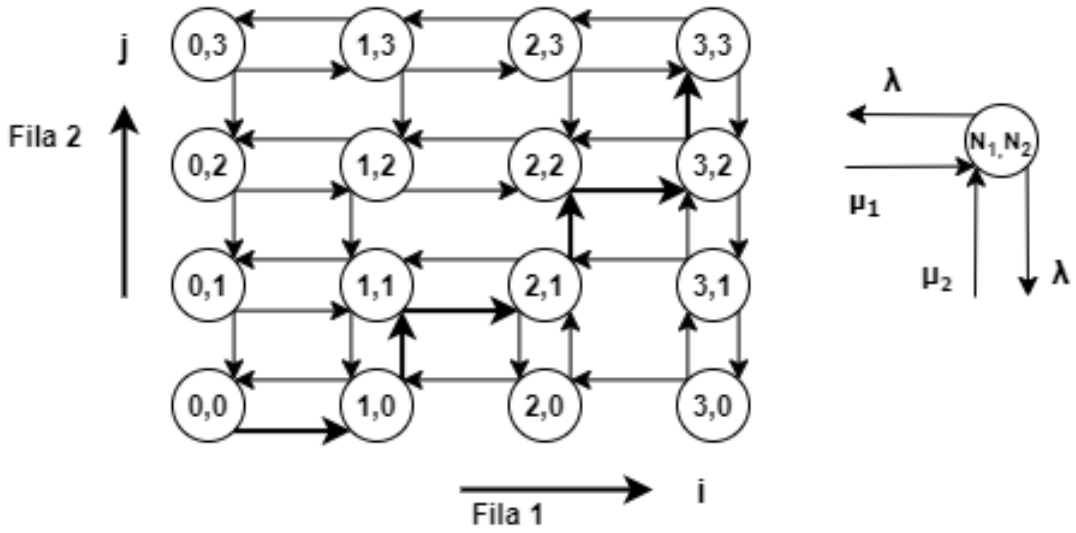



Figura 3.3 – Diagrama de transição de estados para um sistema com dois processadores heterogêneos (CHOW; KOHLER, 1979).

Considere o sistema mostrado na Figura 3.3. Este é constituído por duas filas e dois servidores. O número máximo de tarefas na fila 1 e 2 é definido pelo comprimento $N_1$ e $N_2$. As taxas de serviço dos servidores das filas 1 e 2 são dadas, respectivamente, por $\mu_1$ e $\mu_2$ . A taxa de chegada de tarefas ao sistema é definida pelo parâmetro $\lambda$. Cada nó desse diagrama representa um estado do sistema. As setas destacadas representam a rota definida pelo algoritmo de roteamento utilizado. Os estados conectados pelas setas são chamados de estados balanceados. Os estados acima da rota, definida pelas setas, são chamados de estados superiores desbalanceados e os estados abaixo da rota são chamados de estados inferiores desbalanceados.

O método de solução recursiva para sistemas de dois processadores começa resolvendo as probabilidades de estado pré-normalizadas dos estados de fronteira externa ao longo da coluna $N_1$ ou linha $N_2$, como pode ser visto na Figura 3.3, em termos de algum valor inicial arbitrário atribuído ao estado $(N_1, N_2)$. A Equação de Balanço Global (EBG) utilizada no algoritmo define que, para cada estado de uma rede de filas em equilíbrio, o somatório dos fluxos que saem de um estado é igual ao somatório dos fluxos que chegam a esse estado (BOLCH et al., 2006). Esta conservação do fluxo em estado estacionário pode ser escrita como:

$$\sum_{j \neq i} \pi_i P_{ij} = \sum_{j \neq i} \pi_j P_{ji}. \tag{3.13}$$

A EBG para o estado $(N_1, N_2)$ é então encontrada e usada para resolver a probabilidade pré-normalizada do estado $anterior(N_1, N_2)$, onde a função $anterior(i, j)$ mapeia o estado balanceado $(i, j)$ em seu estado equilibrado anterior na linha de política. Por exemplo, $anterior(N_1, N_2) = (N_1, N_2 - 1)$ na Figura 3.3. O método procede recursivamente,

resolvendo os estados ao longo de uma coluna ou uma linha até $anterior(i, j) = (0, 0)$. Todas as probabilidades previamente calculadas são então normalizadas. Este método recursivo é descrito pelo **Algoritmo 1**.

**Algoritmo 3:** Método recursivo

```
1  Faça (i, j) = (N1, N2), onde N1 é o número máximo de tarefas na fila 1, N2 o
     número máximo de tarefas na fila 2, e (i, j) é o estado atual definido pelo
     algoritmo de roteamento;
2  Faça Pij = Constante (algum valor arbitrário);
3  para K = N1 + N2 até 1 faça
4      (i', j') = anterior (i, j), onde (i', j') é o estado anterior;
5      se i = i' então
6          P0j = X, onde X é uma variável que representa uma probabilidade de
             estado limite temporariamente desconhecida;
7          para I = 1 até i − 1 faça
8              Calcular EBG para o estado (I − 1, j) para encontrar P_Ij;
9          fim
10         Calcular EBG para o estado (i − 1, j), para encontrar X;
11     fim
12     se j = j' então
13         Pi0 = X;
14         para J = 1 até j − 1 faça
15             Calcular EBG para o estado (i, J − 1), para encontrar Pij;
16         fim
17         Calcular EBG para o estado (i, j − 1), para encontrar X;
18     fim
19     Calcular EBG para o estado (i, j), para encontrar Pi'j';
20     (i, j) = (i', j');
21 fim
22 Normalizar todo Pij.
```

Conforme definido pelo algoritmo recursivo, inicialmente os índices $(i, j)$, que representam as filas 1 e 2 do sistema representado na Figura 3.3, são inicializados com ($N_1$, $N_2$). Assim, para o caso em exemplo, onde $N_1 = N_2 = 3$ tem-se os índices $(i, j) = (3, 3)$ e $(i', j') = (3, 2)$. Os valores de $(i, j)$ e $(i', j')$ são atualizados a cada iteração do algoritmo recursivo. $P_{ij}$ é a probabilidade de o sistema estar, em regime permanente (estacionário), no estado $(i, j)$. Para encontrar o valor TT, cujo cálculo depende do valor de $P_{ij}$ para todos os valores de $i$ e $j$ utiliza-se o algoritmo recursivo com o objetivo de calcular a matriz

$P_{ij}$, representada a seguir:

$$Pij = \begin{bmatrix} P_{03} & P_{13} & P_{23} & P_{33} \\ P_{02} & P_{12} & P_{22} & P_{32} \\ P_{01} & P_{11} & P_{21} & P_{31} \\ P_{00} & P_{10} & P_{20} & P_{03} \end{bmatrix}. \tag{3.14}$$

Conforme o algoritmo recursivo, o valor de $P_{33}$ pode ser definido por uma constante com valor arbitrário. Fazendo $P_{33} = 1$, e $i = i'$, os valores de $P_{03}$, $P_{13}$ e $P_{23}$ são calculados recursivamente. A recursão baseia-se em escrever as equações de balanceamento global (EBGs) para cada uma dessas variáveis e fazer substituições sucessivas até obter seus valores numéricos. Esse problema de recursão pode ser resolvido por meio da solução numérica do sistema linear definido pelas EBGs, apresentadas a seguir:

$$\begin{aligned} P_{03}(\lambda + \mu_1) &= P_{13}\mu_1, \\ P_{13}(\lambda + \mu_1 + \mu_2) &= P_{03}\lambda + P_{23}\mu_1, \\ P_{23}(\lambda + \mu_1 + \mu_2) &= P_{13}\lambda + P_{33}\mu_1. \end{aligned} \tag{3.15}$$

Visto que o valor de $P_{33}$ foi inicialmente definido, têm-se um sistema com três equações e três incógnitas. Esse sistema de equações lineares pode ser escrito em uma forma matricial, do tipo $Ax = y$, onde $A$ é uma matriz tridiagonal com os valores das taxas de transições entre os estados, $x$ é o vetor de incógnitas e $y$ é um vetor com valores obtidos através da recursão, ou seja, valores obtidos em iterações anteriores do algoritmo recursivo. Assim, tem-se:

$$\begin{bmatrix} \lambda + \mu_2 & -\mu 1 & 0 \\ -\lambda & \lambda + \mu_1 + \mu_2 & \mu_1 \\ 0 & -\lambda & \lambda + \mu_1 + \mu_2 \end{bmatrix} \begin{bmatrix} P_{03} \\ P_{13} \\ P_{23} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \mu_1 P_{33} \end{bmatrix}. \tag{3.16}$$

Depois de obtidos os valores de $P_{03}$, $P_{13}$ e $P_{23}$, escreve-se a EBG de $P_{ij}$ para obter o valor de $P_{i'j'}$, ou seja, escreve-se a EBG de $P_{33}$ para obter o valor de $P_{32}$ e o valor de $(i, j)$ e $(i', j')$ e uma nova iteração do algoritmo é realizada até que todas as probabilidades sejam calculadas. Um sistema linear geral para o cálculo das probabilidades pode ser escrito como:

$$\begin{bmatrix} \lambda + \mu_2 & -\mu 1 & 0 & 0 \\ -\lambda & \lambda + \mu_1 + \mu_2 & \ddots & 0 \\ 0 & \ddots & \ddots & -\mu_1 \\ 0 & 0 & -\lambda & \lambda + \mu_1 + \mu_2 \end{bmatrix} \begin{bmatrix} P_{0j} \\ P_{1j} \\ \vdots \\ P_{(i-1)j} \end{bmatrix} = \begin{bmatrix} \mu_2 P_{0(j+1)} \\ \mu_2 P_{1(j+1)} \\ \vdots \\ \mu_1 P_{ij} + \mu_2 P(i-1)(j+1) \end{bmatrix}, \quad i = i', \tag{3.17}$$

$$\begin{bmatrix} \lambda+\mu_2 & -\mu 1 & 0 & 0 \\ -\lambda & \lambda+\mu_1+\mu_2 & \ddots & 0 \\ 0 & \ddots & \ddots & -\mu_1 \\ 0 & 0 & -\lambda & \lambda+\mu_1+\mu_2 \end{bmatrix} \begin{bmatrix} P_{i0} \\ P_{i1} \\ \vdots \\ P_{i(j-1)} \end{bmatrix} = \begin{bmatrix} \mu_2 P_{(i+1)0} \\ \mu_2 P_{(i+1)1} \\ \vdots \\ \mu_1 P_(i+1)(j-1) + \mu_2 P_{ij} \end{bmatrix}, \quad j = j'. \tag{3.18}$$

Obtida todas as EBGs para os estados do sistema representado na Figura 3.3 pode-se encontrar os valores de $P_{ij}$. Fazendo $\lambda = 1$, $\mu_1 = 1{,}2$ e $\mu_2 = 1$ obtém-se os seguintes valores para a matriz $P_{ij}$:

$$P_{ij} = \begin{bmatrix} 0,000874 & 0,001457 & 0,003157 & 0.007205 \\ 0,008847 & 0,014017 & 0,028791 & 0,012693 \\ 0,081515 & 0,128486 & 0,059725 & 0,004623 \\ 0,374260 & 0,243954 & 0,028294 & 0,002102 \end{bmatrix}. \tag{3.19}$$

### 3.3.4 Tempo Médio de Permanência no Sistema

Foram utilizados conceitos de teoria de filas e cadeias de Markov para calcular tempo médio de permanência no sistema (TT). Sabe-se que:

$$\pi_i(k) = \sum_{k=1}^{\infty} \pi(k_1, ..., k_n), \tag{3.20}$$

onde $k$ é o número de tarefas na $i$-ésima fila e $\pi_i(k)$ representa a probabilidade de estado em regime permanente relativo ao valor $k$. O número médio de tarefas na $i$-ésima fila é obtido como se segue:

$$K_i = \sum_{k=1}^{\infty} k * \pi_i(k). \tag{3.21}$$

O valor de TT para a $i$-ésima fila é então calculado (BOLCH et al., 2006):

$$TT_i = k_i/\lambda_i, \tag{3.22}$$

onde $\lambda_i$ é a taxa de chegada para a $i$-ésima fila. Para o caso em análise, sistema heterogêneo com duas filas e dois servidores, onde a matriz de probabilidades em regime estacionário é obtida, tem-se:

$$K = \sum_{l=0}^{N_2} \sum_{c=0}^{N_1} (l_i + c_o) * P_{l_i c_o}, \tag{3.23}$$

onde $l_i$ e $c_o$ são, respectivamente, os índices que representam as linhas e as colunas da matriz de probabilidades $P_{ij}$ .

## 3.4 Política com Enfileiramento Baseado em Limiar

Nesta seção é apresentado um modelo de controle dinâmico de carga que utiliza uma política de enfileiramento baseado em um limiar proposto por (LIN; KUMAR, 1984),(LARSEN; AGRAWALA, 1983). Essa política de controle dinâmico é definida para agendar clientes que chegam ao sistema e podem descritos por meio de processos Poisson em um conjunto de dois servidores heterogêneos, cujas taxas de serviço são distintas, mas ambas podem ser modeladas por uma distribuição exponencial, onde, $\mu_1 \neq \mu_2$. O modelo possui um servidor rápido e outro lento. O servidor mais lento é invocado em resposta ao carregamento instantâneo do sistema, medido pelo comprimento da fila de clientes em espera. Em uma política de enfileiramento baseado em um limiar, um comprimento específico da fila é identificado como um limite, além do qual o servidor mais lento é chamado. O servidor mais lento permanece ocupado até concluir o serviço em um cliente e o comprimento da fila é menor que o limite de chamada (LARSEN; AGRAWALA, 1983).

### 3.4.1 Enfileiramento

Considere o sistema de enfileiramento composto por dois servidores visto na Figura 3.4. Supõe-se que o processo de chegada ao nó de origem seja Poissoniano com taxa de chegada $\lambda$ e que os tempos de serviço sejam distribuídos exponencialmente com taxa média $\mu_i$, $(i = 1, 2)$. Sem perda de generalidade, assume-se $\mu_2 \leq \mu_1$ . Para garantir a estabilidade, também deve-se assumir que $\lambda < \mu_2 + \mu_1$. Uma disciplina de fila baseado em um limiar é descrita da seguinte maneira:

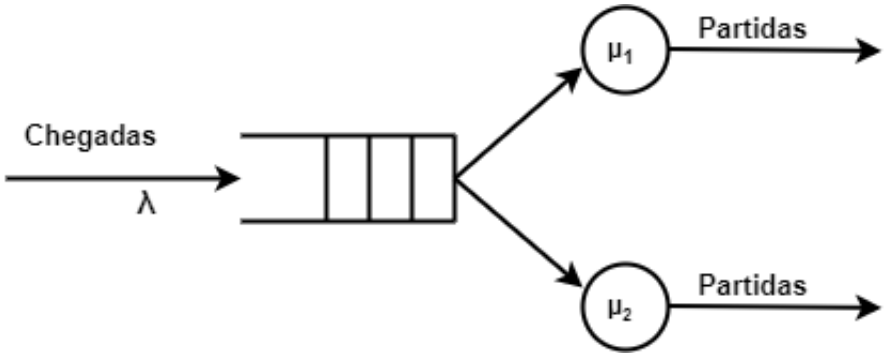



Figura 3.4 – Sistema de enfileiramento.

Disciplina da fila: quando chega uma tarefa ela pode encontrar o seguinte:

1) Ambos os servidores ociosos: a tarefa é despachada para o servidor 1, desde que o servidor 1 ofereça um serviço mais rápido, em média.

2) Ambos os servidores ocupados: a tarefa está na fila no *buffer*.

3) Servidor 1 ocupado e servidor 2 ocioso: a tarefa está na fila no *buffer*. Se o número de clientes no *buffer* exceder $L$, uma tarefa será despachada para o servidor 2. Todos os clientes em espera que chegaram após essa tarefa avançam uma posição. O número inteiro positivo $L$ é chamado de limite do sistema.

Assim, sempre que o servidor 2 estiver disponível, ele começará a servir uma nova tarefa se, e somente se, o número de clientes em espera exceder $L$. Quando o servidor 1 estiver disponível, ele será iniciado servindo a primeira tarefa no *buffer*, independentemente do número total de tarefas em espera. Observe que a disciplina acima mantém o servidor mais rápido ocupado sempre que possível. Foi demonstrado que esta é uma condição necessária para a política ótima (LIN; KUMAR, 1984). O atraso esperado na fila é obtido em termos dos parâmetros do sistema $\lambda$, $\mu_1$, $\mu_2$ e o valor limite $L$.

### 3.4.2 Tempo Médio de Permanência no Sistema

Uma expressão explícita para o tempo médio de permanência no sistema ($TT$) é obtida usando a técnica apresentada em (NI; HWANG, 1985), (KRISHNAMOORTHI, 1963), (ILIADIS; LIEN, 1988).

Seja $s = (s_0, s_1, s_2)$ o estado do sistema de filas onde:

$s_0$ = número de clientes no *buffer*,

$s_j$= 1 ou 0, dependendo de o servidor $j$ estar ocupado ou ocioso.

Esse sistema composto por dois servidores pode ser modelado por um processo de nascimento e morte. Um processo de nascimento e morte é o caso especial de um processo de Markov, no qual as transições do estado $s_j$ são permitidas apenas para os estados vizinhos $s_{j+1}$ e $s_{j-1}$ (KLEINROCK, 1975).

O diagrama de transição de estados correspondente está representado na Figura 3.5. Primeiro, deriva-se o número esperado de clientes no sistema, depois calcula-se o tempo médio do sistema. No diagrama nota-se que a primeira tarefa é despachada para o servidor mais rápido com uma taxa de serviço $\mu_1$. O servidor mais lento permanece inativo enquanto o tamanho da fila com clientes em espera no *buffer* for igual a um valor limiar pré-determinado $L - 1$. Quando o tamanho da fila excede $L$, e o servidor mais rápido estiver ocupado, as tarefas são despachadas para o servidor mais lento com taxa de serviço $\mu_2$. Se o número de clientes no *buffer* for inferior a $L$ quando o período de serviço do servidor mais lento terminar, o servidor mais lento torna-se ocioso.

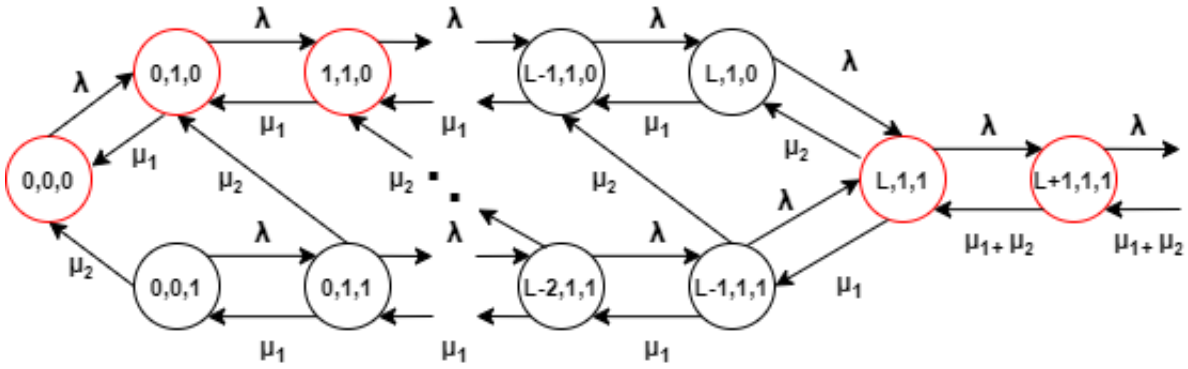



Figura 3.5 – Diagrama de transição de estado resultante de uma política do baseado em um limiar com limite $L$ adaptada de (LIN; KUMAR, 1984).

Para melhor compreensão da política de roteamento, observe os estados em destaque na Figura 3.5. O primeiro cliente que chega ao sistema no estado $s = (0, 0, 0)$ com uma taxa $\lambda$ é atendido pelo servidor mais rápido com uma taxa de serviço $\mu_1$. No estado $s = (0, 1, 0)$ o servidor mais rápido está ocupado enquanto que, o servidor mais lento está ocioso. No estado $s = (1, 1, 0)$ há um cliente na fila de espera e apenas o servidor mais rápido está ocupado. Se os próximos clientes que chegam ao sistema encontrarem o servidor mais rápido ocupado e a fila de espera com o limite $L$ pré-determinado excedido, eles são então atendidos pelo servidor mais lento com uma taxa $\mu_2$, como pode ser observado nos estados $s = (L, 1, 1)$ e $s = (L + 1, 1, 1)$. Se o número de clientes na fila de espera diminuir para um valor inferior a $L$ quando o servidor mais lento atender todos os clientes, ele então torna-se ocioso até que o número de clientes exceda novamente o limite $L$.

Seja $p(s)$ a probabilidade de estado estacionário do estado $s$. A intensidade do tráfego ou o fator de utilização $\rho$ é definido como:

$$\rho = \frac{\lambda}{\mu_1 + \mu_2}. \tag{3.24}$$

Enquanto $\rho < 1$, o sistema terá probabilidades de estado estacionário para $s$. As probabilidades de estado em equilíbrio podem ser expressas como:

$$p(k, 1, 0) = p_1(db^k - c_1\eta_1^{k+1} - c_2\eta_2^{k+1}) \qquad 0 \leq k \leq L, \tag{3.25}$$

$$p(0, 0, 0) = p_1(\frac{d}{b} - 1), \tag{3.26}$$

$$p(k, 1, 1) = p_1(c_1\eta_1^{k+1} + c_2\eta_2^{k+1}) \qquad 0 \leq k \leq L, \tag{3.27}$$

$$p(k,1,1) = \rho^{k-L} p(L,1,1) \qquad k > L, \tag{3.28}$$

onde:

$$\eta_1 = \frac{\lambda + \mu_1 + \mu_2 - \sqrt{(\lambda + \mu_1 + \mu_2)^2 - 4\lambda\mu_1}}{2\mu_1}, \tag{3.29a}$$

$$\eta_2 = \frac{\lambda + \mu_1 + \mu_2 + \sqrt{(\lambda + \mu_1 + \mu_2)^2 - 4\lambda\mu_1}}{2\mu_1}, \tag{3.29b}$$

$$d = \frac{c_1\eta_1^{L+2} + c_2\eta_2^{L+2}}{b^{L+1}}, \quad p_1 = p(0,0,1), \quad b = \frac{\lambda}{\mu_1}, \tag{3.29c}$$

$$c_1 = \frac{1-\eta_1}{\eta_2 - \eta_1}, \quad c_2 = \frac{\eta_2 - 1}{\eta_2 - \eta_1}, \tag{3.29d}$$

e:

$$p_1^{-1} = \begin{cases} \sum_{j=1}^{2} c_j \eta_j^{L+1} \left\{ \frac{\eta_j}{1-b} b^{-(L+2)} + \left(\frac{\rho}{1-\rho} - \frac{\eta_j}{1-b}\right) \right\} & \lambda \neq \mu_1 \\ \\ \sum_{j=1}^{2} c_j \eta_j^{L+1} \left\{ (L+2)\eta_j + \frac{\rho}{1-\rho} \right\} & \lambda = \mu_1. \end{cases} \tag{3.30}$$

O número esperado de clientes no sistema é obtido por:

$$K[n] = p(0,0,1) + \sum_{k=0}^{L}(k+1)p(k,1,0) + \sum_{k=0}^{\infty}(k+2)p(k,1,1). \tag{3.31}$$

Pela fórmula de Little (KLEINROCK, 1975) tem-se que o $TT$ é dado por:

$$\begin{aligned} K[n] &= \lambda * TT, \\ TT &= \frac{K[n]}{\lambda}, \end{aligned} \tag{3.32}$$

o $TT$ é então obtido:

$$TT = \begin{cases} \frac{p_1}{\lambda} \sum_{j=1}^{2} c_j \eta_j^{L+1} \left\{ \left[ \frac{\rho}{1-\rho} - \frac{\eta_j}{1-b} \right] (L+1) + \frac{1}{(1-\rho)^2} + \frac{\eta_j}{(1-b)^2} (b^{-(L+1)} - 1) - \frac{1}{1-\eta_j} \right\}, & \lambda \neq \mu_1 \\ \\ \frac{p_1}{\lambda} \sum_{j=1}^{2} c_j \eta_j^{L+1} \left\{ \frac{\eta_j}{2} (L+1)^2 + (\frac{\rho}{1-\rho} - \frac{\eta_j}{2})(L+1) + \frac{1}{(1-\rho)^2} - \frac{1}{1-\eta_j} \right\}, & \lambda = \mu_1. \end{cases} \tag{3.33}$$

O $TT$ expressa o tempo médio de permanência no sistema, considerando o atraso médio das tarefas na fila, o comprimento da fila fixado pelo limite $L$, a taxa chegadas de tarefas no sistema e as taxas de serviços de ambos os servidores.

## 3.5 Política de Junção à Fila Mais Curta

Em sistemas distribuídos e redes de comunicação, a política de junção à fila mais curta (JSQ) tem sido usada como mecanismo básico de balanceamento de carga ou roteamento de tarefas. Ela tem várias aplicações nos processos de planejamento de capacidade, dimensionamento do sistema, detecção de congestionamentos, estudo do desempenho de diferentes arquiteturas de multiprocessadores, entre outras (LIN; RAGHAVENDRA, 1996). No entanto, uma grande desvantagem da política JSQ é que, quando aplicada a um sistema que consiste em um grande número de servidores, ela necessita das informações de estado de todos os servidores do sistema para tomar decisões de atribuição de tarefas. O uso de algoritmos dinâmicos randomizados é uma forma de evitar a exigência de informações sobre todas as ocupações dos servidores (MUKHOPADHYAY; MAZUMDAR, 2015).

O sistema de enfileiramento considerado pela política JSQ é modelado por um sistema heterogêneo de múltiplos processadores como mostrado na Figura 3.6. O sistema é composto por 2 servidores, com tempo de processamento exponencial e taxa média de atendimento $\mu_i$, assim o processo de saída de tarefas do sistema segue um processo de Poisson. O tempo entre as chegadas de tarefas também é caracterizado por uma distribuição exponencial com taxa média $\lambda$. A disciplina seguida pelas filas é do tipo FCFS (*First-Come-First-Served*). Para garantir a estabilidade deve-se assumir que $\lambda < \mu_2 + \mu_1$.

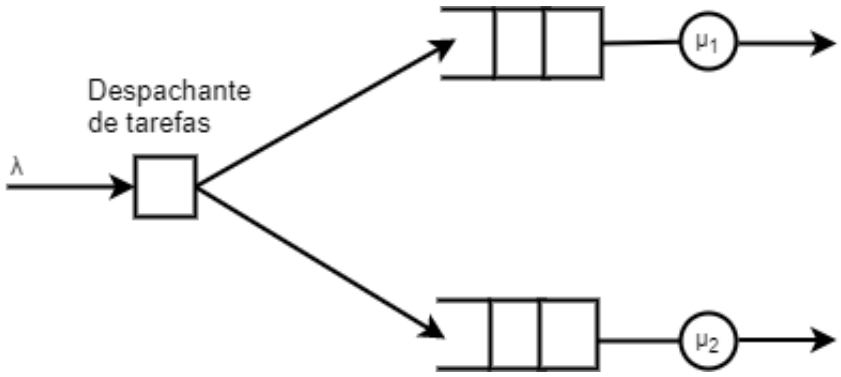



Figura 3.6 – Sistema heterogêneo com a política JSQ.

Um despachante de tarefas é utilizado para atribuir tarefas às filas de espera. No instante da chegada a tarefa é atribuída à fila com o menor número de tarefas. Sejam $n_1$ e $n_2$ o número de tarefas na fila 1 e 2, respectivamente. A política de roteamento JSQ pode ser declarada da seguinte maneira:

- se $n_1 < n_2$, atribui uma tarefa que chega ao sistema à fila 1, e à fila 2 caso contrário;
- se o $n_1 = n_2$, atribui a próxima tarefa à fila com maior taxa de serviço $\mu_i$, $i = (1, 2)$.

Observe que quando o número de tarefas é o mesmo para as duas filas a próxima tarefa que chega ao sistema é atribuída ao servidor mais rápido, garantindo um menor tempo de espera para as demais tarefas que chegaram ao sistema. A política JSQ é baseada nas informações de estado do sistema para melhorar o desempenho, uma vez que sempre será necessário conhecer a quantidade de tarefas nas filas de espera. A política JSQ é modelada aqui como um processo de Markov bidimensional sobre os estados s($n_1$, $n_2$). Na Figura 3.7 é apresentado o diagrama de transição de estados para $\mu_1 = 2$ e $\mu_2 = 1$.

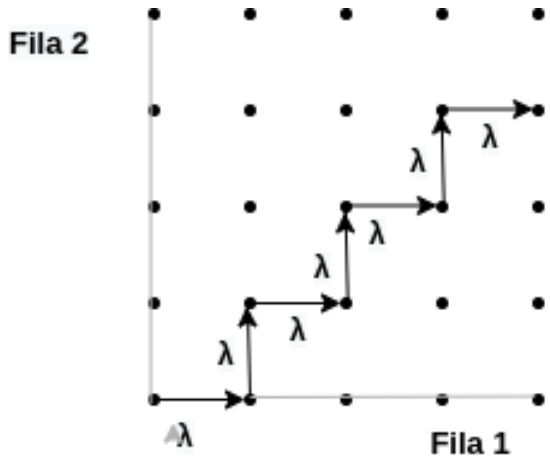



Figura 3.7 – Diagrama de transição de estados para um sistema com dois processadores heterogêneos onde $\mu_1 = 2$ e $\mu_2 = 1$.

As probabilidades de estado em equilíbrio, necessárias para se calcular o tempo total de permanência das tarefas no sistema ($TT$), foram obtidas utilizando a técnica recursiva apresentada na seção 3.3.3. Após obtida a matriz de probabilidades o $TT$ é então calculado como mostrado na seção 3.3.4.

# 4 Balanceamento de Carga Aplicado nas Maternidades

Nesta seção, é descrito o modelo de simulação utilizado. São mostrados os resultados das simulações do balanceamento de carga aplicado ao sistema perinatal. Por fim, os resultados obtidos são analisados, comparados e discutidos.

## 4.1 Modelo de Simulação

A rede perinatal do cenário proposto apresentada na Figura 2.1 pode ser estendida para Figura 4.1. As taxas de chegadas de gestantes às maternidades são definidas pelos parâmetros $\lambda_1$ e $\lambda_2$, onde a taxa total de chegada ao sistema é $\lambda$=$\lambda_1$+$\lambda_2$. As taxas de atendimento nas maternidades $M_1$ e $M_2$ são definidas pelos parâmetros $\mu_1$ e $\mu_2$. A tomada de decisões é baseada na carga de chegada das gestantes às maternidades. Ao chegar em uma das duas maternidades, no processo de atendimento e admissão, é verificada a disponibilidade dos serviços utilizando os algoritmos de balanceamento de carga, caso a gestante seja admitida na maternidade em que teve o primeiro atendimento, ela é direcionada para os cuidados internos da maternidade. No caso de rejeição, a gestante é direcionada para a segunda maternidade.

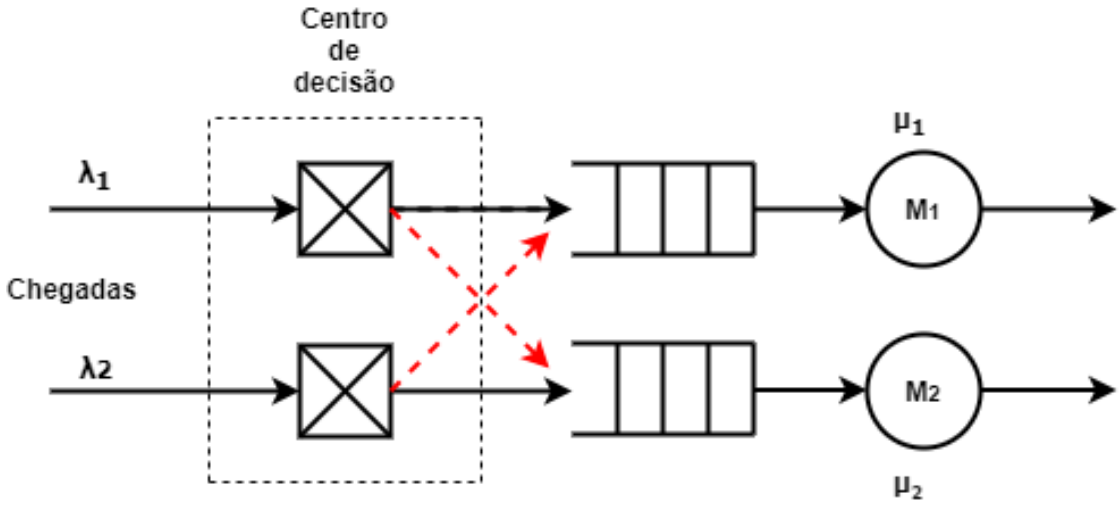



Figura 4.1 – Rede perinatal composta por duas maternidades.

Os algoritmos de balanceamento de carga foram aplicados ao sistema perinatal com base em duas maternidades, levando em conta o processo de chegada, admissão e rejeição de gestantes a ambas as maternidades, o que é representado pela Figura 4.1. Além disso, com o objetivo de modelar o processo de admissão nas etapas internas de uma

maternidade, conforme ilustrado na Figura 2.2, aplicou-se dados estatísticos obtidos das duas maternidades para escalonar as gestantes entre CO e CC. Entende-se que após a admissão de uma gestante em uma maternidade, há algumas possíveis salas de atendimento onde ela pode ser encaminhada, e isso também exige equilíbrio de carga. Por exemplo, quer seja necessário que a gestante vá para a UTI, o algoritmo é aplicado com base na carga necessária e nos recursos das maternidades.

Considerando dados coletados de duas maternidades anônimas de Goiânia, pode-se obter a taxa média de chegadas de gestantes por dia baseado no número de atendimentos e procedimentos realizados dentro de certo período. Na Tabela 4.1 são apresentados as taxas médias de atendimentos, de partos normais e de cesáreas nas duas maternidades.

Tabela 4.1 – Número médio estimado de procedimentos por dia.

| **Tipo de serviço** | **Maternidade 1** | **Maternidade 2** |
|---|---|---|
| Atendimentos | 31,4 | 14,2 |
| Partos normais | 7,9 | 5,4 |
| Partos cirúrgicos (cesáreas) | 5,3 | 3,2 |

Os dados obtidos relativos às taxas de chegadas, aos atendimentos e aos procedimentos realizados nas duas maternidades foram utilizados como entrada para as simulações da rede perinatal. As taxas de serviços foram obtidas de acordo com a quantidade de leitos de internação, leitos para partos e leitos de UTI disponíveis em cada uma das maternidades.

## 4.2 Resultados e Discussões

Alguns cenários são analisados e os resultados discutidos e comparados. A Tabela 4.2 apresenta os algoritmos e os parâmetros do sistema utilizados para as simulações.

Tabela 4.2 – Parâmetros do Sistema.

| **Símbolo** | **Descrição** |
|---|---|
| $\mu_i$ | Taxa de serviço das maternidades |
| $\lambda$ | Taxa de chegadas de gestantes |
| $TT$ | Média de permanência na maternidade |
| $PR$ | Roteamento proporcional |
| $R$ | Política de tempo mínimo de resposta |
| $T$ | Política de tempo mínimo do sistema |
| $TP$ | Política de máxima vazão |
| $L$ | Política com enfileiramento baseado em Limiar |
| $JSQ$ | Política de junção à fila mais curta |
| $HBB-LB$ | Política inspirada no comportamento das abelhas |
| $M_i$ | Maternidades |

### 4.2.1 Primeiro Cenário

Neste cenário foi utilizado os dados reais informados na Tabela 4.1, exceto que, nenhuma das maternidades possuem UTIs para mães, o que foi incluído apenas para as simulações. Para esta simulação foi possível considerar que:

- $M_1$ possui uma quantidade maior de leitos em relação a $M_2$. Portanto, $\mu_1 > \mu_2$;
- Ambas possuem Centros de Obstetrícia;
- Ambas possuem Centro Cirúrgico;
- $M_1$ e $M_2$ possuem UTIs para as mães.

Na primeira etapa foram aplicados os algoritmos de roteamento para as gestantes que chegam ao sistema perinatal, a fim de decidir em qual maternidade, $M_1$ ou $M_2$, elas serão admitidas. A Figura 4.2 apresenta a curva de TT para $\lambda_1 = 31$ e $\lambda_2 = 14$ , onde $\lambda_1$ e $\lambda_2$ são as taxas de chegadas de gestantes por dia em $M_1$ e $M_2$, respectivamente. Como definido $\mu_1 > \mu_2$, sendo que $\mu_i$ é a taxa de atendimentos em um dia, de acordo com a disponibilidade dos recursos nas duas maternidades.

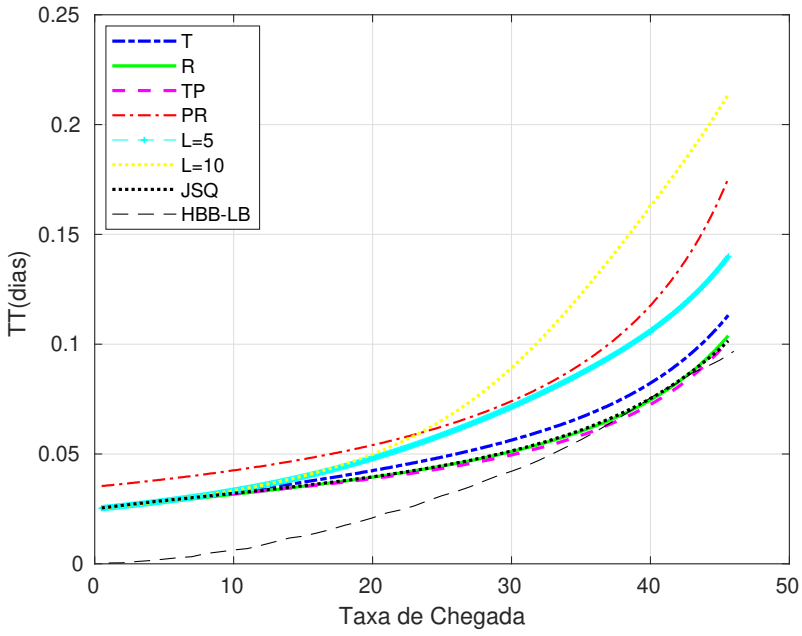
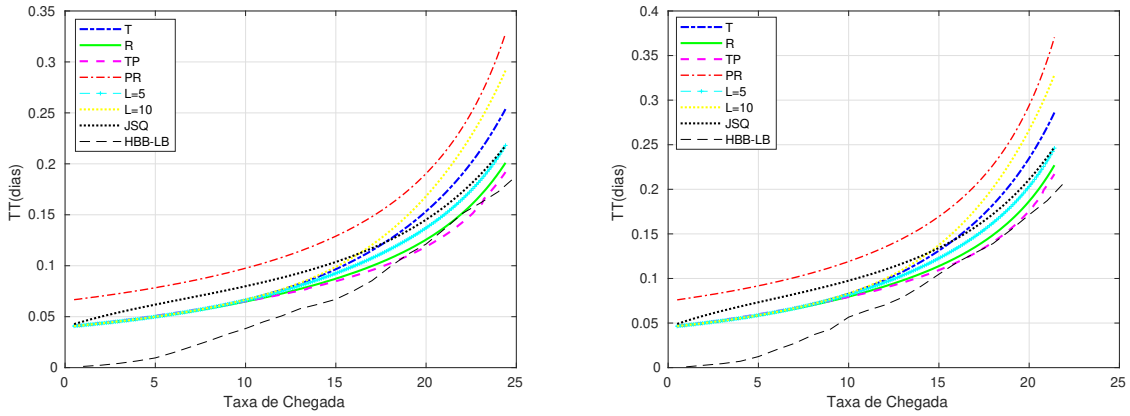



Figura 4.2 – TT para os algoritmos de roteamento PR, R, T, TP, L=5, L=10, JSQ e HBB-LB com $\mu_1 = 40$ e $\mu_2 = 17$ para admissão a uma das maternidades

Uma vez concluída a admissão das gestantes em uma das duas maternidades, utilizou-se o algoritmo em cada uma individualmente, de acordo com a carga de gestantes que necessitavam ir para o CC ou para o CO. Admitiu-se que, seja para a fila de CO

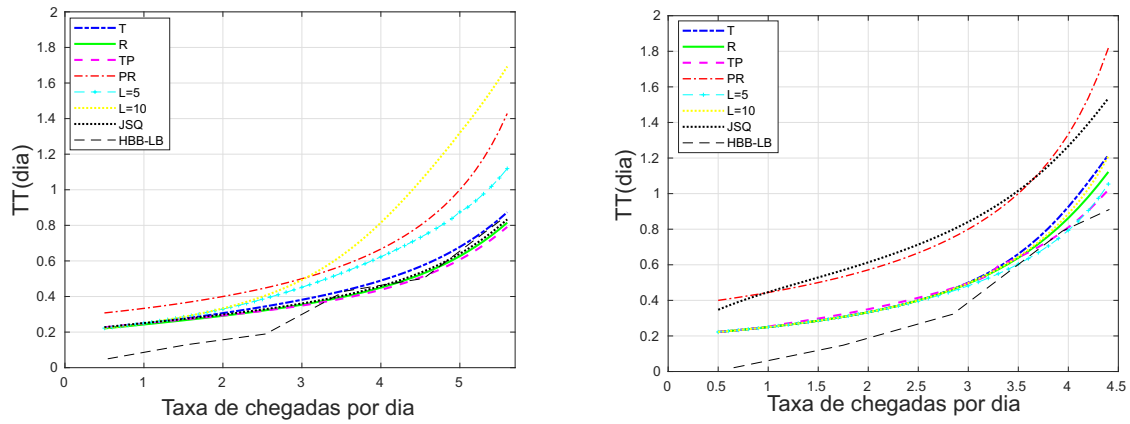
ou CC, sempre haverá uma gestante em espera aguardando parto normal ou cesárea. A Figura 4.3 apresenta as curvas de TT para $M_1$ e $M_2$ no balanceamento de carga entre CO e CC.




(a) $M_1$ com $\mu_1 = 25$ e $\mu_2 = 5{,}5$ (b) $M_2$ com $\mu_1 = 22$ e $\mu_2 = 4{,}8$

Figura 4.3 – TT para os algoritmos de roteamento PR, R, T, TP, L=5, L=10, JSQ e HBB-LB para CO e CC.

As duas maternidades possuem UTIs para as mães, portanto, as taxas de serviços $\mu_1$ e $\mu_2$ estão de acordo com a carga exigida. As taxas de serviços para as UTIs são baixas, dado que, o tempo de permanência nessas alas são altos. A Figura 4.4 apresenta as curvas de TT para admissão às UTIs.




(a) $M_1$ com $\mu_1 = 5$ e $\mu_2 = 2$ (b) $M_2$ com $\mu_1 = 5$ e $\mu_2 = 0{,}5$

Figura 4.4 – TT para os algoritmos de roteamento PR, R, T, TP, L=5, L=10, JSQ e HBB-LB para as UTIs.

### 4.2.2 Segundo Cenário

Para este cenário, os dados relativos ao tempo de chegadas, os tempos de serviços e capacidade foram escolhidas considerando valores menores que os dados reais, levando em

consideração carga de gestantes que entram no sistema moderada. Considerou-se também que:

- $M_1$ possui uma quantidade maior de leitos em relação a $M_2$. Portanto, $\mu_1 > \mu_2$;
- Ambas possuem Centros de Obstetrícia;
- Ambas possuem Centro Cirúrgico;
- $M_1$ e $M_2$ possuem UTIs para as mães.

Na primeira etapa, de admissão a uma das duas maternidades, como definido $\mu_1 > \mu_2$, onde $M_1$ possui uma quantidade maior de leitos que $M_2$. A Figura 4.5 apresenta a curva de TT para $\lambda_1 = 20$ e $\lambda_2 = 12$.

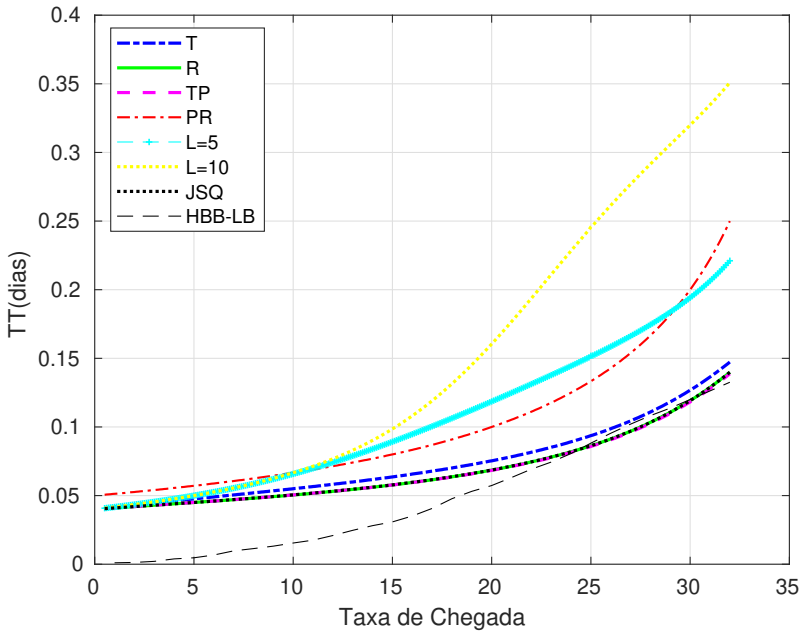
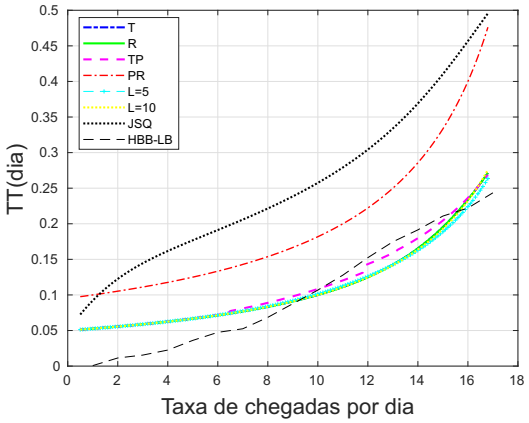
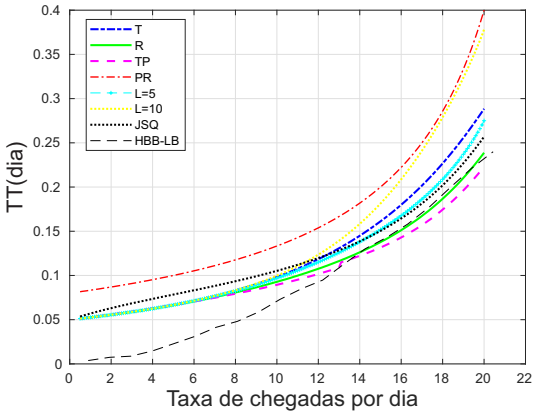



Figura 4.5 – TT para os algoritmos de roteamento PR, R, T, TP, L=5, L=10, JSQ e HBB-LB com $\mu_1 = 25$ e $\mu_2 = 15$ para admissão a uma das maternidades.

Na segunda etapa considerou-se que em $M_1$ a capacidade oferecida para partos no CC é baixa, portanto o valor de $\mu_2$ é relativamente pequeno. $M_2$ possui CO e CC então os valores de $\mu_1$ e $\mu_2$ estão de acordo com cenário de simulação proposto. A Figura 4.6 apresenta as curvas de TT obtidas para esta etapa.



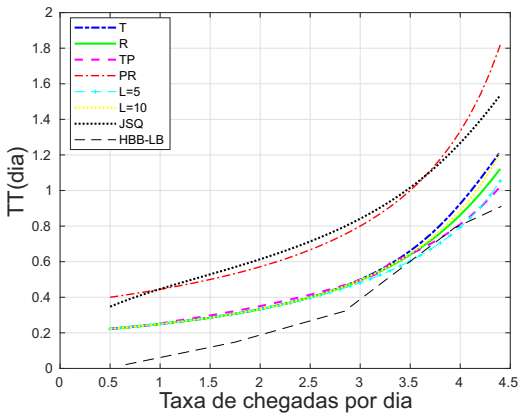

(a) $M_1$ com $\mu_1 = 20$ e $\mu_2 = 1$



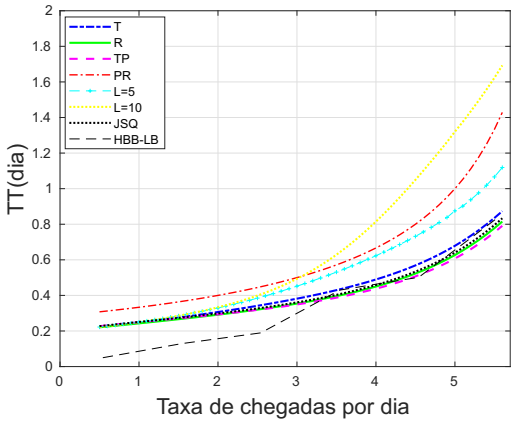

(b) $M_2$ com $\mu_1 = 20$ e $\mu_2 = 5$

Figura 4.6 – TT para os algoritmos de roteamento PR, R, T, TP, L=5, L=10, JSQ e HBB-LB para CO e CC.

Assim como no primeiro cenário, considerou-se que as duas maternidades possuem UTIs para as mães. A Figura 4.7 apresenta as curvas de TT para admissão às UTIs.




(a) $M_1$ com $\mu_1 = 5$ e $\mu_2 = 0.5$




(b) $M_2$ com $\mu_1 = 5$ e $\mu_2 = 2$

Figura 4.7 – TT para os algoritmos de roteamento PR, R, T, TP, L=5, L=10 para as UTIs.

Analisando os resultados obtidos observa-se que, dentre os algoritmos de roteamento PR, R, T, TP, L = 5, L = 10 e JSQ o TP é o que apresenta menor TT em todas as etapas simuladas para o sistema perinatal. A eficiência desse algoritmo é mais evidente à medida que $\lambda$ aumenta. Isso se deve ao fato de que o algoritmo TP considera o efeito da taxa de chegada. O sistema alcança um desempenho melhor sob condições de carga pesada quando a quantidade de processamento total é utilizada. Os algoritmos R, T e JSQ levam em consideração apenas as taxas de serviços das maternidades ou a quantidade de gestantes na fila não conhecendo, portanto, o comportamento das taxas de chegadas. O

algoritmo HBB-LB obteve resultados melhores que o algoritmo TP conseguindo reduzir ainda mais o tempo de permanência das gestantes nas maternidades. Esse algoritmo atualiza as informações da capacidade de cada maternidade ao longo do tempo, sendo possível prever quando uma maternidade estará sobrecarregada. Quando há a sobrecarga de uma das maternidades, é feito o balanceamento das gestantes que estão na fila de espera sendo possível melhorar o tempo de resposta dos serviços oferecidos. Portanto, o algoritmo HBB-LB provou ser eficiente, diminuindo o tempo de permanência das gestantes na rede.

Quanto à política de controle baseado em um limiar, o TT cresce à medida que os valores das taxas de serviço $\mu_1$ e $\mu_2$ das duas maternidades se aproximam, e o comprimento da fila definida por um limite $L$ aumenta. Essa política torna-se eficaz quando a maternidade com atendimento mais rápido possui taxas de serviços relativamente maiores que as taxas do servidor lento e o limite da fila $L$ não for alto. A maternidade com atendimento lento só receberá a gestante quando o limite da fila $L$ exceder, portanto, quanto maior for esse limite $L$, a tendência é que, as gestantes fiquem mais tempo na fila de espera até serem atendidas.

## 4.3 Tempo Médio de Permanência na Maternidade

O tempo médio de permanência representa o período médio em dias que as gestantes permanecem internadas na maternidade. Muitos fatores estão associados ao intervalo de tempo de internação, como o tipo do parto e a classificação de risco. No SUS observou-se, para o período de 2008 a 2012, uma média de permanência para partos normais de 2,0 dias e para partos cesáreos de 2,6 dias (SAúDE, 2012). As médias de permanência para as gestações de alto risco foram mais elevadas, situando-se em 3,2 dias para partos normais e 4,2 dias para partos cesáreos. O tempo médio de permanência considerando ambos os partos, em gestações de baixo e alto risco, foi de 2,3 dias nesse período. Segundo os dados informados pelas duas maternidades anônimas, o tempo de permanência das gestantes no período analisado foi de 3,5 dias para a maternidade $M_1$ e 3 dias para a maternidade $M_2$. Esses valores se apresentam maiores do que o tempo médio de permanência observado no país. Os tempos médios de permanência obtidos para o primeiro cenário são apresentados na Tabela 4.3.

Tabela 4.3 – Tempo médio de permanência para o primeiro cenário

| **Algoritmo** | **Maternidade 1** | **Maternidade 2** | **Média do Sistema** |
|---|---|---|---|
| Real | 3,5 | 3 | 3,25 |
| PR | 1,65 | 2,36 | 2,0 |
| R | 1,16 | 1,53 | 1,34 |
| T | 1,29 | 1,93 | 1,56 |
| TP | 1,13 | 1,40 | 1,26 |
| L5 | 1,47 | 1,61 | 1,54 |
| L10 | 2,19 | 1,75 | 1,97 |
| JSQ | 1,15 | 1,88 | 1,51 |
| **HBB-LB** | **1,10** | **1,34** | **1,22** |

O tempo médio de permanência foi entre 1,22 e 2,0 dias, valores estes que se apresentaram na maioria dos algoritmos utilizados, melhores que os dados reais apresentados pelo SUS. Os tempos médios de permanência obtidos para o segundo cenário são apresentados na Tabela 4.4.

Tabela 4.4 – Tempo médio de permanência para o segundo cenário

| **Algoritmo** | **Maternidade 1** | **Maternidade 2** | **Média do Sistema** |
|---|---|---|---|
| Real | 3,5 | 3 | 3,25 |
| PR | 2,53 | 2,08 | 2,30 |
| R | 1,52 | 1,35 | 1,43 |
| T | 1,64 | 1,32 | 1,48 |
| TP | 1,40 | 1,16 | 1,28 |
| L5 | 1,54 | 1,61 | 1,57 |
| L10 | 1,84 | 2,42 | 2,13 |
| JSQ | 1,93 | 1,46 | 1,69 |
| **HBB-LB** | **1,38** | **1,15** | **1,26** |

Para o segundo cenário a média de permanência obtida situou-se entre 1,26 dias a 2,3 dias. Analisando os resultados obtidos é possível considerar que foram satisfatórios, pois são compatíveis com os dados reais e mostram o quão importante pode ser um algoritmo de balanceamento de carga. O algoritmo HBB-LB foi um dos algoritmos que apresentou melhores valores do que os dados reais, conseguindo reduzir o tempo de permanência das gestantes nas maternidades. O tempo médio total do sistema foi menor que os outros algoritmos, sendo mais eficiente à medida que as taxas de chegada aumentam.

Portanto, em um sistema perinatal que enfrenta problemas de alocação e atendimento de gestantes exatamente devido às altas taxas de chegadas e uma quantidade limitada de recursos para atender a demanda, modelos de roteamento podem ser aplicados obtendo-se resultados satisfatórios, já que as taxas de chegadas $\lambda$ são tratadas a fim de melhorar o desempenho do sistema diminuindo o tempo em que as gestantes permanecem

na rede. Diminuir o tempo de permanência em qualquer unidade hospitalar é algo essencial. Os gastos com internações representam uma grande fatia dos gastos com assistência à saúde no Brasil. Vale ressaltar ainda, que, quanto ao número de dias de internação, nem sempre mais significa melhor. Uma melhor gerência da média de permanência é fundamental para manter o bom desempenho de uma unidade hospitalar. Em relação às maternidades, diminuir esse indicador contribui para manter o equilíbrio financeiro, a qualidade da assistência e a segurança das gestantes, além da otimização dos processos internos e a redução do índice de infecções e perdas. Os algoritmos de balanceamento de carga se mostraram estratégias interessantes para diminuir esse tempo de internação das gestantes em maternidades.

## 4.4 Simulação de Eventos Discretos

Nesta seção é apresentado um modelo de simulação de eventos discretos com o objetivo de consolidar os resultados obtidos pelos algoritmos de balanceamento de carga. É feito uma análise do tempo nas filas de espera das unidades perinatais e do tempo total de permanência das gestantes na rede perinatal. Os resultados obtidos para o tempo de permanência são comparados com os resultados obtidos pelos algoritmos de balanceamento de carga descritos na seção 4.

### 4.4.1 Caracterização do Modelo

A sequência lógica de funcionamento da rede perinatal aqui simulada é representada nas Figuras 2.1 e 2.2 descritas na seção 2. Na primeira etapa, quando as gestantes chegam ao sistema é feito o escalonamento destas a uma das duas maternidades como é representado na Figura 4.8. Nesta etapa da simulação, as gestantes são direcionadas à uma das duas maternidades de acordo com a disponibilidade de leitos e recursos, o que é feito proporcionalmente às suas capacidades. O tempo entre as chegadas de gestantes obedece a uma distribuição exponencial.

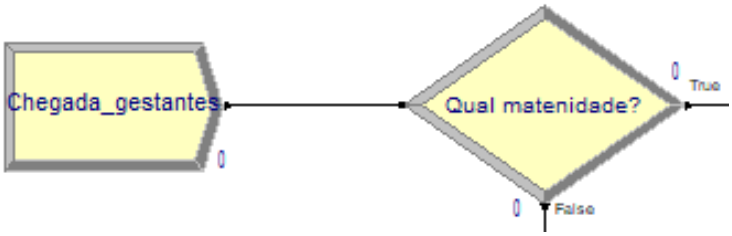



Figura 4.8 – Escalonamento de gestantes a uma das duas maternidades.

Após a admissão das gestantes à uma das duas maternidades como pode ser observado na Figura 2.2 elas são encaminhadas para os leitos pré-parto. No entanto, antes

disso, elas passam por uma análise de risco onde é verificado o quadro clínico e a urgência do parto. Quando o parto é considerado de emergência as gestantes são direcionadas para o CC ou CO de acordo com risco do parto. Considerou-se que o tempo médio de espera na recepção é de 30 minutos. Além disso, o tempo médio para realização do parto normal dura em média 1 hora, e do parto cirúrgico dura em média 2 horas. Foi estabelecido que, 10% das gestantes que são admitidas à uma das duas maternidades são submetidas ao parto de emergência onde, em 70% dos casos são realizados partos cesáreos e 30% são partos normais. As gestantes que não são submetidas ao parto de emergência passam por uma avaliação do quadro clínico antes de serem encaminhadas para o CO ou CC. A média de realização de partos normais em hospitais e maternidades públicas está em torno de 60% a 70% o que não é a taxa idealizada pela comunidade médica internacional. Diante disso, foi definido na simulação a proporção dos partos realizados nas duas maternidades onde 60% dos partos são normais e 30% são cesáreas. A Figura 4.9 apresenta o modelo simulado nessa etapa para as duas maternidades.

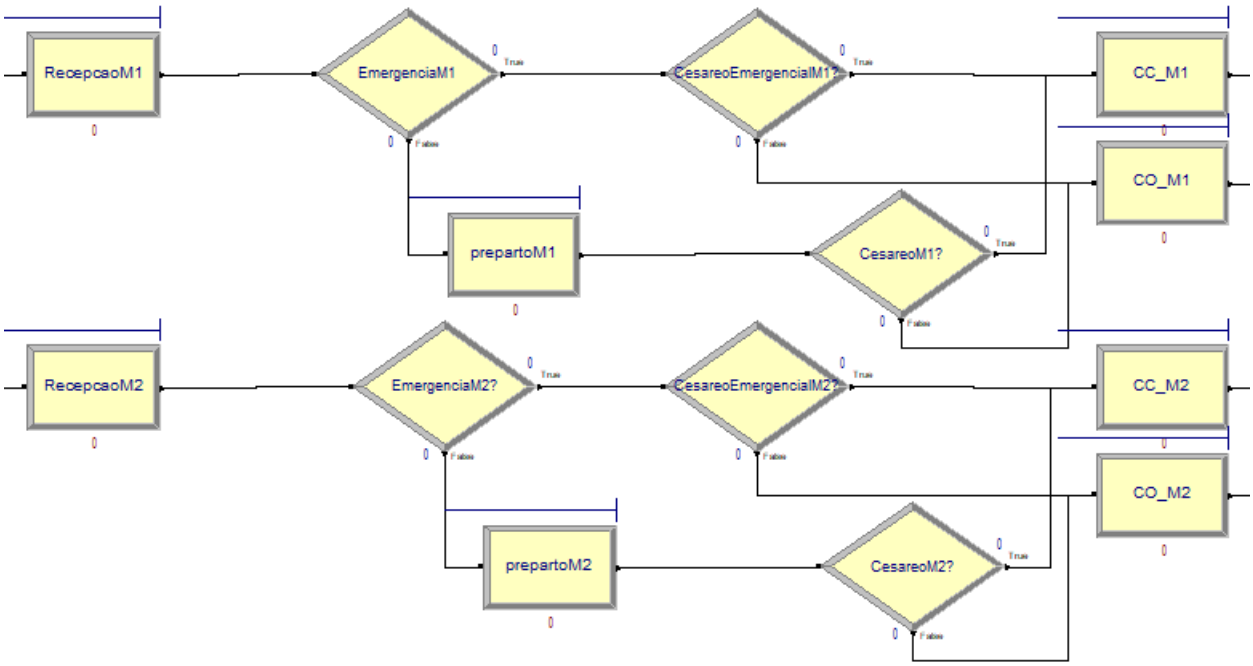



Figura 4.9 – Escalonamento interno de gestantes a ao CO ou CC.

Após o parto nos casos em que há complicações no com risco de vida para a puérpera esta pode ser deslocada para a UTI onde permanece até que se recupere totalmente. Caso ela não necessite de UTI ela é encaminhada ao leito pós-parto até o momento da alta. Aqui, considerou-se que em 5% dos casos as puérperas são encaminhadas para UTI, onde permanecem em média 10 dias. Caso não haja complicações as puérperas são encaminhadas para os leitos pós-parto onde permanecem em observação. Se o parto foi normal a puérpera deve permanecer em média 24 horas em observação e 48 horas em caso de parto cesáreo

antes de recebem alta da maternidade. A Figura 4.10 apresenta o modelo de simulado para as duas maternidades.

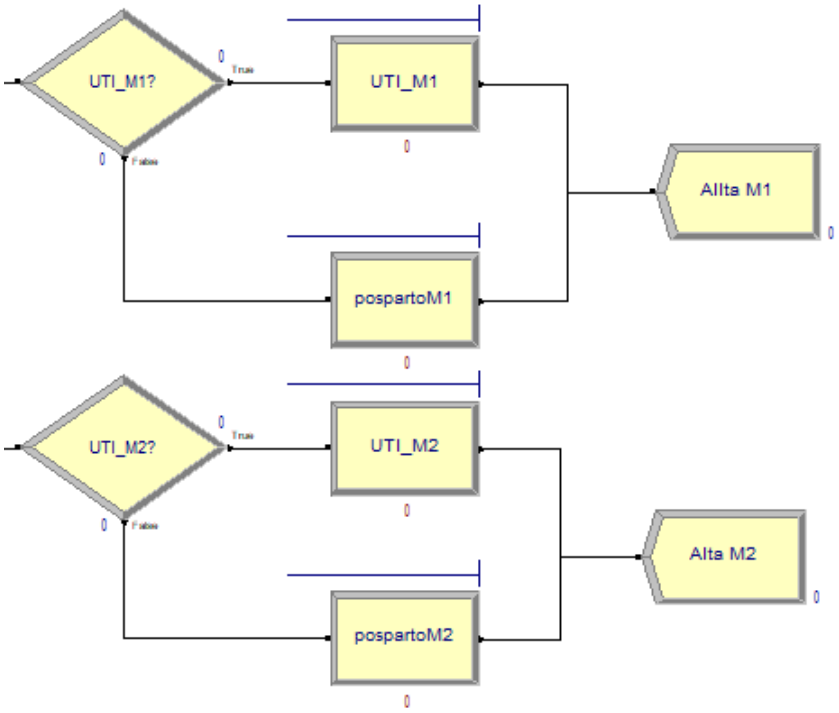



Figura 4.10 – Escalonamento de gestantes alta ou UTI.

O modelo de simulação completo com todas as etapas percorridas pelas gestantes à partir de sua admissão até sua alta é descrito na Figura 4.11. Foram simulados dois cenários descritos nas seções 4.2.1 e 4.2.2 e o modelo utilizado foi o mesmo, apenas substituindo as variáveis de entrada, como a taxa de chegadas de gestante e a capacidade de cada maternidade.

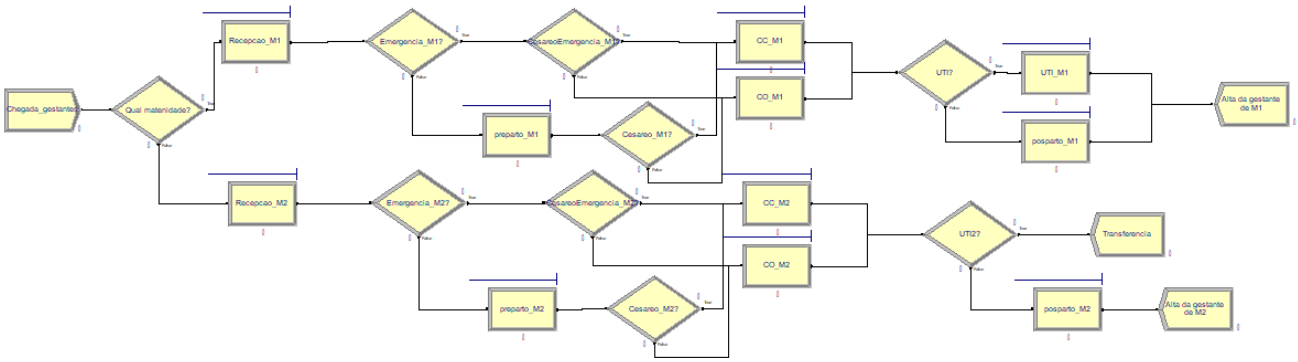


Figura 4.11 – Modelo completo simulado.

Foram utilizados os dados de entrada de gestantes apresentados na Tabela 4.1. Quanto aos parâmetros introduzidos no modelo tem-se que a maternidade $M_1$ possui

47 leitos obstétricos disponíveis, tanto para partos normais quanto para cesáreas). A maternidade $M_2$ possui 30 leitos. A Tabela 4.5 apresenta os recursos disponíveis nas maternidades diariamente utilizados na simulação.

Tabela 4.5 – Número de leitos e pessoal por dia.

| **Descrição** | **Maternidade 1** | **Maternidade 2** |
|---|---|---|
| Leitos obstétricos ( parto normal e cesárea) | 47 | 30 |
| Enfermeiros(as) e Técnicos(as) | 15 | 10 |
| Médico Obstetra | 5 | 5 |
| Médico Pediatra | 4 | 4 |

O modelo de simulação foi construído no software Rockwell Arena 15.1 em sua versão para estudantes. Os resultados são testados em um computador com processador de 1,7 GHz e memória RAM de 4GB. Para simular o sistema real com precisão, alguns parâmetros importantes de simulação devem ser determinados, tais como número de réplicas, duração da replicação e período de aquecimento. Além disso, a determinação de intervalos de confiança para as variáveis de desempenho é um componente fundamental no processo de análise de resultados. Para isso, é possível calcular a quantidade de replicações que garantam um intervalo de confiança ideal usando a seguinte fórmula:

$$\overline{X}(n) \pm t_{n-1,1-\alpha/2}\sqrt{\frac{S^2(n)}{n}}, \tag{4.1}$$

onde $S_2(n)$ é o desvio padrão estimado e $\overline{X}(n)$ é a média estimada para $X_1, X_2, ..., X_n$ valores de um indicador para $n$ replicações. O valor $t_{n-1,1-\alpha/2}$ é o ponto crítico da distribuição $t$ com $n-1$ graus de liberdade e uma cobertura igual a $\frac{1-\alpha}{2}$. Para garantir um intervalo de confiança de 95% para os indicadores de desempenho são necessárias 20 replicações.

A fim de determinar a duração da simulação e o período de aquecimento, foi observado o intervalo de tempo em que os dados foram obtidos. O modelo de simulação pode alcançar rapidamente sua condição de estado estável, pois as entidades do sistema não crescem exponencialmente no tempo. Portanto, o modelo de simulação é executado com 20 réplicas, com uma duração de execução de 3 anos, sendo que o primeiro teve como objetivo apenas preencher o sistema, sendo assim considerado período de aquecimento.

### 4.4.2 Resultados Computacionais

Nesta seção, são apresentados os resultados do modelo de simulação de eventos discretos e comparados aos resultados obtidos pelos algoritmos de balanceamento de carga. Foram obtidos os tempos médios de espera em cada setor das maternidades. A Tabela 4.6 apresenta o tempo médio de permanência das gestantes nas filas dado em horas para o primeiro cenário como descrito na seção 4.2.1.

Tabela 4.6 – Tempo médio de espera nas filas em horas.

| **Fila** | **Maternidade 1** | **Maternidade 2** |
|---|---|---|
| Leitos pré-parto | 0,0013 | 0,0004 |
| CC | 2,64 | 4,74 |
| CO | 0,64 | 1,08 |
| Leitos pós-parto | 0,41 | 0,70 |
| UTI | 6,96 | 7,49 |

O tempo de espera na maioria das filas se apresentaram baixos, porém, para CC o tempo de espera foi considerado elevado dado que as gestantes que são submetidas aos partos cirúrgicos demandam maior urgência por se tratar de partos de alto risco. O tempo de espera também foi considerado elevado para as puérperas que aguardam leitos de UTI, já que em muitos casos a transferência é de extrema urgência. Deve-se ressaltar que não foi modelado o falecimento de puérperas nessas filas, o que certamente pode ocorrer em casos de longa espera. Também não foi modelada a possibilidade do parto ser transferido para outra maternidade, o que possivelmente impactaria no tempo médio de espera.

Também foram feitas simulações para o cenário descrito na seção 4.2.2. A Tabela 4.7 apresenta o tempo médio de permanência das gestantes nas filas dado em horas.

Tabela 4.7 – Tempo médio de espera nas filas em horas.

| **Fila** | **Maternidade 1** | **Maternidade 2** |
|---|---|---|
| Leitos pré-parto | 0,0006 | 0,00009 |
| CC | 2,62 | 4,70 |
| CO | 0,59 | 1,17 |
| Leitos pós-parto | 0,39 | 0,79 |
| UTI | 1,54 | 0,12 |

O tempo de espera para UTI no segundo cenário é menor pois a taxa de gestantes que são atendidas pela maternidade $M_2$ é inferior em relação a $M_1$ dado que sua capacidade física instalada é menor e a taxa de puérperas que necessitam de terapia intensiva pequena. Quanto aos tempo de espera para CC, este ainda foi considerado alto para $M_1$. O tempo de espera conforme a categoria de risco é um indicador fundamental para eficiência da maternidade. A redução dos tempos de espera, especialmente em casos graves, pode melhorar a qualidade da assistência às gestantes, proporcionando menores índices de complicações na hora do parto. Longos tempos de espera contribuem para a superlotação das maternidades, que podem levar a uma série de problemas, incluindo sofrimento para as gestantes que esperam por atendimento, recusa de novas gestantes, alta tensão para a equipe assistencial e óbitos maternos e neonatais.

Também foram obtidos o tempo médio de permanência das gestantes na rede

perinatal para os dois cenários analisados. A Tabela 4.8 apresenta a comparação entre os tempos obtidos pela simulação e pelos algoritmos de balanceamento de carga.

Tabela 4.8 – Tempo médio de permanência simulado.

| **Algoritmos** | **Cenário 1** | **Cenário 2** |
|---|---|---|
| Simulado no Arena | 2,21 | 2,7 |
| PR | 2,0 | 2,30 |
| R | 1,34 | 1,43 |
| T | 1,56 | 1,48 |
| TP | 1,26 | 1,28 |
| L5 | 1,54 | 1,57 |
| L10 | 1,97 | 2,13 |
| JSQ | 1,51 | 1,69 |
| **HBB-LB** | **1,22** | **1,26** |

O tempo médio de permanência obtido pelo modelo de simulação esteve próximo do tempo médio real observado pelo SUS que é de 2,3 dias. No entanto, o tempo médio de permanência real estimado apresentado pelas duas maternidades, $M_1 = 3{,}5$ dias e $M_2 = 3$ dias, no período analisado, é maior que o tempo obtido nas simulações. Alguns fatores contribuem para o tempo de permanência mais elevado, como a qualidade de serviço oferecida e a quantidade de recursos disponibilizados pelas unidades. Há também que se considerar o quadro clínico das gestantes que chegam a essas unidades, já que não é uma variável determinística o que causa impacto na assitência e consequentemente no tempo em que elas devem permanecer no sistema.

Observou ainda, que os tempos de permanência obtidos pelos algoritmos de balanceamento foram menores que os obtidos nas simulações. Nas simulações de eventos discretos o critério para admissão à uma das duas maternidades foi baseado apenas na capacidade de $M_1$ e $M_2$, e o encaminhamento das gestantes foi feito proporcionalmente a quantidade de leitos e recursos disponíveis, enquanto que, nos algoritmos de balanceamento de carga esse encaminhamento foi feito com base em políticas de roteamento. As políticas de roteamento conseguem aplicar com maior eficiência o encaminhamento das gestantes que entram no sistema o que diminui o tempo em que elas permanecem no sistema.

# 5 Dimensionamento de Recursos nas Maternidades

O aumento do número de nascimentos, a diminuição dos leitos obstétricos e a consequente superlotação nas unidades perinatais levanta questões relativas ao planejamento, previsão e o cálculo do tamanho adequado para estas unidades. A indisponibilidade de leitos obstétricos, de leitos de UTI e de uma estrutura mínima que ofereça assistência às parturientes pode levar à rejeição de internação dessas pacientes e a necessidade de transferências para outras maternidades. Uma gestante rejeitada de uma unidade perinatal pode transbordar para outra unidade na mesma rede. Por conseguinte, uma maternidade pode receber tanto as suas próprias pacientes como as pacientes que transbordam de outras maternidades.

Como a demanda por serviços nas unidades perinatais aumenta significativamente, apresenta-se como alternativa aumentar o número de leitos ou adequar a demanda ao número de leitos disponíveis, o que muitas vezes só é obtido à custa de baixas taxas de ocupação (JONES, 2012). Diante da diminuição dos leitos obstétricos a melhor opção seria adequar o número de partos e atendimentos à capacidade física da unidade perinatal. Geralmente, algumas maternidades elaboram, ao início de cada ano, um relatório de produção estimada para cada mês buscando regular o número de partos que poderão ser realizados, o número de procedimentos cirúrgicos e o número de atendimentos e demais serviços oferecidos. No entanto, as taxas de superlotação estão aumentando devido ao mal planejamento da disponibilização dos recursos existentes.

As decisões sobre o planejamento da capacidade nos hospitais são tomadas geralmente sem o aporte de análises quantitativas baseadas em estudos e modelos (LAFFEL; BLUMENTHAL, 1989). O que explica a diminuição do número de leitos obstétricos e o aumento nas taxas de ocupação das unidades perinatais. A variabilidade no tempo de permanência das gestantes nas unidades perinatais causam um grande impacto no funcionamento diário dessas unidades e nos requisitos de capacidade. Se esta variabilidade for desconsiderada durante a modelagem, surgirá uma representação irrealista e estática da realidade (BRUIN et al., 2007).

O planejamento e o dimensionamento dos recursos em unidades hospitalares são amplamente abordados em diversos trabalhos sob a perspectiva da teoria de filas (GREEN et al., 2006), (BRUIN et al., 2007), (TSEYTLIN, 2009) (BRUIN et al., 2010), (ZAIED, 2010), (YANKOVIC; GREEN, 2011), (VÉRICOURT; JENNINGS, 2011), (PEHLIVAN et al., 2012), (MANDELBAUM; MOMČILOVIĆ; TSEYTLIN, 2012), (PEHLIVAN; AUGUSTO; XIE, 2014), (PEHLIVAN, 2014), (ARMONY et al., 2015). Diversos autores

já estudaram o modelo de filas de espera com o uso da equação de *Erlang-B* para o dimensionamento de recursos em unidades hospitalares (GREEN; NGUYEN, 2001), (HARPER; SHAHANI, 2002), (GORUNESCU; MCCLEAN; MILLARD, 2002), (KOKANGUL, 2008), (LI et al., 2009), (BRUIN et al., 2010), (BELCIUG; GORUNESCU, 2015), (ESSEN; HOUDENHOVEN; HURINK, 2015). A equação de *Erlang-B* possui variáveis importantes que possibilitam o cálculo de várias medidas de interesse para o planejamento tanto de recursos humanos como da capacidade física dos serviços hospitalares.

Nesta seção é descrito um modelo matemático de enfileiramento para o cálculo e estimativa da capacidade das unidades perinatais com base no tempo de permanência das gestantes no sistema.

## 5.1 Modelo Matemático

Como já foi introduzido, as pacientes de redes perinatais são urgentes, e, portanto, uma capacidade de atendimento inadequada em uma maternidade faz com que a gestante espere muito ou seja transferida para outra unidade, o que pode causar longas distâncias de deslocamento para outra unidade de saúde resultando em rejeição e transbordamento de gestantes. Esse problema pode e deve ser evitado por meio de um apropriado dimensionamento e operação eficiente dos recursos disponíveis (PEHLIVAN; AUGUSTO; XIE, 2013).

Baseado na teoria de filas muitos estudos analisam modelos de atraso que são aplicados para problemas de planejamento de capacidade em unidades hospitalares (GREEN; NGUYEN, 2001), (GREEN, 2002),(BRUIN et al., 2010). O fluxo dos pacientes que entram em uma fila de espera é analisado para quantificar a probabilidade de rejeição e a probabilidade de atraso, que são importantes medidas de desempenho para um sistema de enfileiramento. O objetivo principal é encontrar soluções para um melhor dimensionamento e adequação dos recursos existentes na unidade hospitalar. Como em uma unidade perinatal a rejeição de gestantes é um problema real a ser evitado, um modelo de enfileiramento adequado é o M/G/$m$/$m$ que incorpora o bloqueio aos clientes que entram no sistema e encontram todos os servidores ocupados.

### 5.1.1 Modelo de Enfileiramento M/G/$m$/$m$

Considere o sistema de enfileiramento M/M/$m$/$m$ descrito na Figura 5.1. Os clientes chegam de acordo com um processo de Poisson. O tempo entre as chegadas obedece a uma distribuição exponencial e os tempos de serviço são exponencialmente distribuídos. Uma fila M/M/$m$/$m$ trata-se então de um sistema com $m$ servidores e capacidade $m$, sem fila de espera. Portanto, se os $m$ servidores estiverem todos ocupados quando um novo cliente chegar, ele será perdido pois o sistema está cheio, ocorrendo então o bloqueio do sistema.

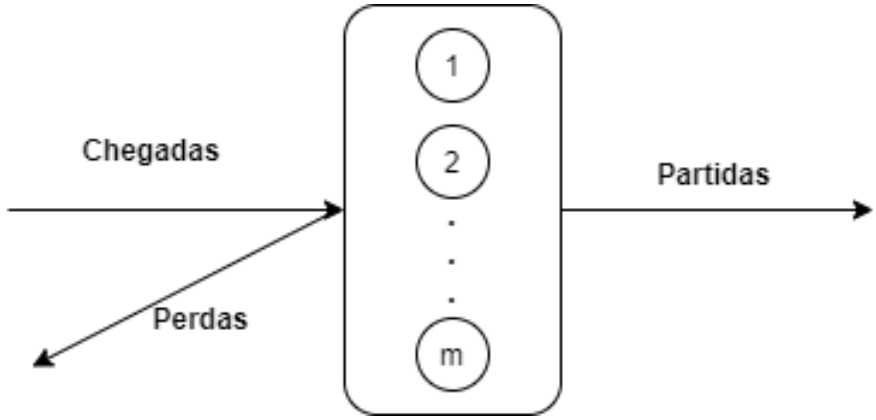



Figura 5.1 – Sistema de enfileiramento M/M/$m$/$m$.

O estado do sistema é definido pelo número de clientes presentes no sistema e o espaço de estado é finito. O estado pode ser descrito por um processo de nascimento e morte para qual o diagrama de transição de estados é mostrado na Figura 5.2.

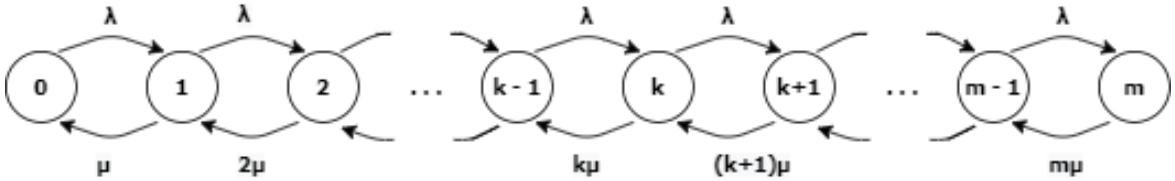



Figura 5.2 – Diagrama de transição estados para um sistema M/M/$m$/$m$.

Seja $p_k$ a probabilidade de haver k clientes no sistema em um momento arbitrário em equilíbrio, onde $k = 0, 1, 2, ..., m$. As probabilidades de estado em equilíbrio para $\{pk; 0 \leq k \leq m\}$ podem ser expressas como:

$$\begin{aligned} \mu p_1 &= \lambda p_0, \\ \lambda p_{k-1} + (k+1)\mu p_{k+1} &= (\lambda + k\mu)p + k, \qquad 1 \leq k \leq m, \\ \lambda p_{m-1} &= m\mu p_m. \end{aligned} \tag{5.1}$$

Isto é equivalente ao conjunto:

$$k\mu p_k = \lambda p_{k-1}, \qquad 1 \leq k \leq m, \tag{5.2}$$

onde:

$$p_k = \frac{\lambda}{k\mu} p_{k-1} = \frac{\lambda^2}{k(k-1)\mu^2} p_{k-2} = ... = \frac{1}{k!}(\frac{\lambda}{\mu})^k p_0, \qquad 1 \leq k \leq m. \tag{5.3}$$

A intensidade do tráfego ou o fator de utilização $\rho$ é definido como:

$$\rho = \frac{\lambda}{\mu}. \tag{5.4}$$

Encontrando $p_0$ através da condição de normalização:

$$\sum_{k=0}^{m} p_k = 1, \tag{5.5}$$

obtém-se a distribuição de probabilidade para o número de clientes no sistema.

$$p_k = p_0 \frac{\rho^n}{m!} = \frac{\rho^m/m!}{\sum_{k=0}^{m} \rho^k/k!}, \qquad 1 \leq k \leq m. \tag{5.6}$$

Esta é a distribuição de Poisson truncada. Os clientes que chegam quando todos os servidores estão ocupados são bloqueados. A probabilidade de haver k clientes no sistema imediatamente antes de uma chegada é igual à probabilidade de haver k clientes no sistema em um tempo arbitrário dado na Equação (5.6) (WOLFF, 1989). Portanto, a probabilidade de bloqueio, também conhecida como equação de *Erlang-B* ou é dada por:

$$P_m = \frac{(\lambda/\mu)^m/m!}{\sum_{k=0}^{m}(\lambda/\mu)^k/k!} \tag{5.7}$$

Pode-se provar que também para uma distribuição geral do tempo de serviço as probabilidades $P_n$ são dadas pela Equação (5.7), onde $\rho = \lambda\mu$ é frequentemente referido como a carga oferecida ao sistema. Obtém-se então a equação de *Erlang-loss* para o modelo M/G/$m$/$m$:

$$P_m = \frac{(\lambda\mu)^m/m!}{\sum_{k=0}^{m}(\lambda\mu)^k/k!} \tag{5.8}$$

### 5.1.2 Descrição do Modelo

Considere o modelo estrutural do fluxo do gestantes em uma maternidade mostrado na Figura 5.3. As gestantes que chegam ao sistema são admitidas se houver leitos disponíveis. Caso a gestante encontre todos os leitos ocupados ela é recusada e deixa o sistema. Na prática, uma admissão recusada pode resultar em um desvio para outra maternidade não-preferível. O número de admissões recusadas pode ser interpretado como um indicador do nível de serviço e é importante para a qualidade do atendimento (BRUIN et al., 2010). As maternidades mantêm registro do número de admissões contabilizada como o número de procedimentos realizados, aqui considerado, como o número de partos. O tempo de permanência é contabilizado desde a entrada da gestante na maternidade até o momento da alta, mesmo que essa seja transferida para outra ala da maternidade. O

tempo de permanência é uma métrica essencial para evitar congestionamento e atrasos nos atendimentos. A capacidade de uma maternidade é medida em termos de leitos operacionais e a quantidade de pessoal disponível é baseada nessa capacidade operacional. Isto é feito através de uma relação de pessoal por leito operacional. A rejeição de gestantes também pode acontecer em outros ambientes da maternidade após a admissão e o parto. Há casos em que as puérperas necessitam de cuidados intensivos devido a complicações durante e ou antes do parto. Isso é um agravante pois, com as gestantes em estado crítico de saúde a transferência para outra unidade que possua leitos para tratamento intensivo se torna algo delicado e não preferível.

Admite-se que as gestantes chegam à maternidade com distribuição de Poisson, pois se mostrou um bom ajuste para as chegadas não programadas. Esta distribuição pode descrever também a maioria dos departamentos clínicos de um hospital.

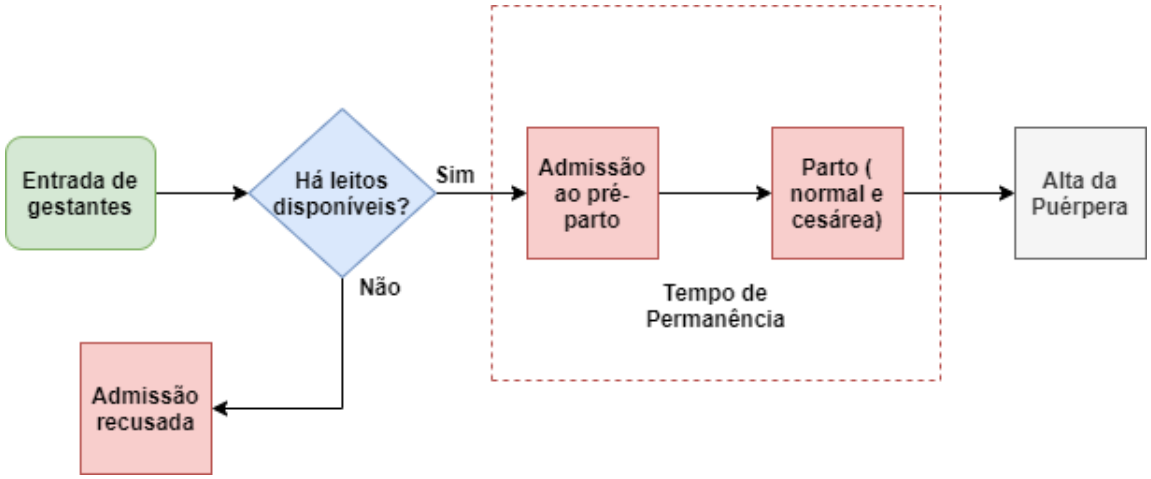



Figura 5.3 – Modelo estrutural do fluxo de gestantes em uma maternidade.

As unidades perinatais exigem um certo número de leitos para admitir e tratar todas as gestantes. O número necessário de leitos depende do tempo de permanência de cada gestante e do número de internações e atendimentos por dia. A taxa de admissões e de perdas são variáveis aleatórias. Caso fossem variáveis determinísticas, ou seja, a cada dia exatamente $\lambda$ gestantes chegassem com um tempo médio de permanência $1/\mu$, então a cada dia haveria exatamente $\lambda/\mu = \rho$ leitos ocupados quando não ocorresse nenhum bloqueio. Entretanto, na prática, o número de gestantes e a média de permanência são variáveis estocásticas, já que o tempo de permanência é diferente para cada gestante e o número de internações varia por dia. Para garantir que haja leitos suficientes em um ambiente estocástico, certamente precisa-se de mais leitos do que $\rho$ para lidar com a flutuação do número de gestantes admitidas. Porém, na prática, nem sempre será possível admitir todas as gestantes, porque então o número de leitos necessários teria que ser extremamente grande e uma grande parte dos leitos ficariam ociosos na maior parte do tempo perdendo eficiência na utilização dos recursos (ESSEN; HOUDENHOVEN; HURINK, 2015).

Com os dados obtidos das duas maternidades anônimas de Goiânia pode-se usar a Equação (5.8) de perda de *Erlang* para determinar a média de permanência e a taxa de rejeição de gestantes, o número de leitos necessários para se atingir uma taxa de rejeição predefinida e a taxa média de admissões desejada para a quantidade de leitos disponíveis. Na Equação (5.8) tem-se que:

- $\lambda$ = número médio chegadas por dia;
- $\mu$ = tempo médio de permanência das gestantes;
- $n$ = Número de leitos disponíveis;
- $P_n$ = taxa de rejeição de gestantes.

Usando a fórmula de Little (KLEINROCK, 1975) a taxa de ocupação é obtida fazendo:

$$T_o = \frac{\lambda\mu(1 - P_n)}{n}. \tag{5.9}$$

O número médio de leitos ocupados pode ser encontrado por:

$$\overline{N} = \lambda\mu(1 - P_n). \tag{5.10}$$

Substituindo $P_n$ na Equação (5.10) usando (5.8), obtém-se a expressão para o número médio de leitos ocupados:

$$\overline{N} = \lambda\mu(1 - \frac{(\lambda\mu)^n/n!}{\sum_{k=0}^{n}(\lambda\mu)^k/k!}). \tag{5.11}$$

O número médio de leitos ocupados pode ser obtido nos dados informados pelas duas maternidades e o número de leitos $m$ e o tempo médio de permanência são variáveis conhecidas.

## 5.2 Resultados e Discussões

Os registros do número de partos ocorridos nas duas maternidades apresentados na Tabela 4.1 foram usados para quantificar o número de admissões e a taxa de ocupação. A taxa de ocupação permite avaliar o grau de utilização dos leitos operacionais na maternidade, medir o perfil de utilização e gestão desses leitos e está diretamente relacionado ao intervalo de substituição e à média de permanência. A taxa de ocupação recomendável situa-se entre 75 e 85%. A taxa de ocupação abaixo de 75% indica baixa utilização e ineficiência na gestão hospitalar. Para as duas maternidades analisadas, a média das taxas de ocupação

situou-se entre 80 e 90%. À Tabela 5.1 apresenta o número de leitos operacionais nas duas maternidades, as médias de permanência observadas e o número de chegadas diárias estimadas no período analisado. A medida que os leitos vão sendo ocupados, o número de leitos operacionais disponíveis é obtido através da taxa de ocupação.

Tabela 5.1 – Número de leitos, média de permanência e número de chegadas diárias.

| **Descrição** | **Maternidade 1** | **Maternidade 2** |
|---|---|---|
| Leitos obstétricos ( parto normal e cesárea) | 47 | 30 |
| Média de Permanência (dias) | 3,5 | 3 |
| Números de chegadas diárias ($\lambda$) | 13,86 | 9,03 |

Com esses dados, o modelo de perda de Erlang pode então ser usado para determinar o número de leitos necessários quando a média de permanência, a taxa de chegadas e a probabilidade de bloqueio são dadas. Assim, um valor aceitável para a probabilidade de bloqueio pode ser escolhida. Considerando os dados apresentados na Tabela 4.1 estima-se a taxa de chegadas nas duas maternidades. Como o número de admissões recusadas não é registrada, considerou-se que 5% das gestantes que chegam à maternidade são rejeitadas. Portanto, a demanda diária estimada de gestantes que chegam às duas maternidades, entre admissões recusadas ou não, foi de 13,86 chegadas por dia para a primeira maternidade e 9,03 chegadas por dia para a segunda maternidade. Vale ressaltar que, a demanda considerada foi apenas para os leitos para partos (normal ou cesáreas).

Foram feitas simulações de diversos cenários, reduzindo a demanda nessas duas maternidades para diferentes taxas de rejeição. O objetivo é encontrar o menor número de leitos para os quais a taxa de bloqueio é menor ou igual a taxa fixada. A Tabela 5.2 apresenta o número de leitos requeridos para diferentes taxas de rejeição. Foram utilizados os tempos médios de permanência reais informado pelas maternidades e os tempos calculados pelos algoritmos de balanceamento de carga para primeiro o cenário da seção 4 apresentados na Tabela 4.3.

Tabela 5.2 – Número de leitos necessários para diferentes taxas de rejeição na Maternidade $M_1$.

| **Número de chegadas diárias ($\lambda$)** | **Tempo médio de permanência (dias)** | **Nº leitos requeridos para rejeição de:** | | | | | |
|---|---|---|---|---|---|---|---|
| | | **25%** | **20%** | **15%** | **10%** | **5%** | **2%** |
| 13,86* | 3,5* | 39 | 42 | 45 | 49 | **54** | 59 |
| | 1,65 (PR) | 19 | 21 | 23 | 25 | **28** | 31 |
| | 1,16 (R) | 14 | 15 | 17 | 18 | **21** | 23 |
| | 1,29 (T) | 15 | 17 | 18 | 20 | **23** | 25 |
| | 1,13 (TP) | 14 | 15 | 16 | 18 | **20** | 22 |
| | 1,47 (L5) | 17 | 19 | 20 | 22 | **25** | 28 |
| | 2,19 (L10) | 25 | 27 | 29 | 32 | **35** | 39 |
| | 1,15 (JSQ) | 14 | 15 | 16 | 18 | **20** | 23 |
| | 1,10 (HBB-LB) | 13 | 14 | 16 | 17 | **20** | 22 |
| 11,1(-20%) | 3,5* | 31 | 34 | 36 | 40 | **44** | 48 |
| | 1,65 (PR) | 16 | 17 | 19 | 20 | **23** | 25 |
| | 1,16 (R) | 11 | 12 | 13 | 15 | **17** | 19 |
| | 1,29 (T) | 12 | 13 | 15 | 16 | **19** | 21 |
| | 1,13 (TP) | 11 | 12 | 13 | 14 | **16** | 19 |
| | 1,47 (L5) | 14 | 15 | 17 | 19 | **21** | 23 |
| | 2,19 (L10) | 20 | 22 | 24 | 26 | **29** | 32 |
| | 1,15 (JSQ) | 11 | 12 | 13 | 15 | **17** | 19 |
| | 1,10 (HBB-LB) | 11 | 12 | 13 | 14 | **16** | 18 |

*Valores reais estimados

O tempo médio permanência das gestantes na maternidade $M_1$ é considerada alta em comparação ao tempo médio de permanência informado pelo SUS, mesmo considerando que o tempo médio de permanência para gestações de alto risco elevem a taxa de permanência global da maternidade. O tempo médio de permanência nacional considerando ambos os partos, em gestações de baixo e alto risco, foi de 2,3 dias em um período de 4 anos (SAúDE, 2012). Dado esse elevado tempo de permanência, para se obter taxas de admissões recusadas menores é necessário aumentar o número de leitos ou que a demanda de gestantes diminua cerca de 20%. No entanto, dificilmente a demanda teria uma diminuição tão significativa.

Para a demanda e a média de permanência reais informadas pela maternidade, a taxa de rejeição é de 12%, o que significa que aproximadamente a cada 9 gestantes que procuram o serviço para parto, uma não encontra leito disponível para internação. Qual a taxa de admissões recusadas é razoável e, portanto, aceitável ainda é alvo de grandes discussões, no entanto, alguns especialistas em políticas hospitalares muitas vezes buscam atingir uma meta de 5% (BRUIN et al., 2010). A Tabela 5.2 revela que, para atingir uma meta de 5% de admissões de gestantes recusadas seriam necessários 54 leitos, ou seja, 7 leitos a mais, dado que a maternidade possui 47 leitos operacionais disponíveis. Caso a

demanda diminua em 20% seria necessário apenas mais um leito para que a meta de 5% seja alcançada.

Analisando os resultados obtidos para os tempos médios de permanência calculados pelos algoritmos de balanceamento de carga, observa-se que, à medida que a média de permanência das gestantes cai o número necessário de leitos operacionais cai significativamente. Os tempos médios de permanência obtidos pelos algoritmos situam-se entre 1,10 a 2,19 dias. Hoje no Brasil a exigência mínima do Ministério da Saúde para permanência da parturiente na maternidade é de 24 horas, isso se, o parto tenha sido normal e não teve complicações antes, durante ou após o parto. Portanto, se o tempo médio se mantivesse em torno do tempo mínimo exigido pelo Ministério da Saúde, o planejamento e a gerência dos recursos nas unidades perinatais seria mais facilmente sustentado, pois poderia haver alocação e transferência de leitos entre unidades perinatais próximas, evitando a superlotação de uma maternidade pequena e a ociosidade em uma maternidade de grande porte.

As simulações também foram feitas para a maternidade $M_2$. A Tabela 5.3 apresenta o número de leitos requeridos para diferentes taxas de rejeição para o primeiro cenário.

Tabela 5.3 – Número de leitos necessários para diferentes taxas de rejeição na Maternidade $M_2$.

| **Número de chegadas diárias ($\lambda$)** | **Tempo médio de permanência (dias)** | **Nº leitos requeridos para rejeição de:** | | | | | |
|---|---|---|---|---|---|---|---|
| | | **25%** | **20%** | **15%** | **10%** | **5%** | **2%** |
| 9,03* | 3* | 22 | 24 | 26 | 29 | **32** | 35 |
| | 2,36 (PR) | 18 | 19 | 21 | 23 | **26** | 29 |
| | 1,53 (R) | 12 | 13 | 14 | 16 | **18** | 20 |
| | 1,93 (T) | 15 | 16 | 18 | 19 | **22** | 24 |
| | 1,40 (TP) | 11 | 12 | 13 | 15 | **17** | 19 |
| | 1,61 (L5) | 13 | 14 | 15 | 17 | **19** | 21 |
| | 1,75 (L10) | 14 | 15 | 16 | 18 | **20** | 23 |
| | 1,88 (JSQ) | 14 | 16 | 17 | 19 | **21** | 24 |
| | 1,34 (HBB-LB) | 11 | 12 | 13 | 14 | **16** | 18 |
| 7,2(-20%) | 3* | 18 | 20 | 21 | 23 | **26** | 29 |
| | 2,36 (PR) | 14 | 16 | 17 | 19 | **21** | 24 |
| | 1,53 (R) | 10 | 11 | 12 | 13 | **15** | 17 |
| | 1,93 (T) | 12 | 13 | 14 | 16 | **18** | 20 |
| | 1,40 (TP) | 9 | 10 | 11 | 12 | **14** | 16 |
| | 1,61 (L5) | 10 | 11 | 12 | 14 | **16** | 18 |
| | 1,75 (L10) | 11 | 12 | 13 | 15 | **17** | 19 |
| | 1,88 (JSQ) | 12 | 13 | 14 | 16 | **18** | 20 |
| | 1,34 (HBB-LB) | 9 | 10 | 11 | 12 | **14** | 15 |

*Valores reais estimados

A maternidade $M_2$ possui uma quantidade menor de leitos em relação a $M_1$, além de apresentar uma demanda menor por serviços de parto. Com o tempo médio de permanência de 3 dias como informado, é possível alcançar a meta de 5% de admissões recusadas aumentando apenas 2 leitos. Esse resultado revela que, a maternidade consegue manter uma taxa de admissões recusadas baixa com os recursos disponíveis sem a necessidade de aumento significativo no número de leitos ou diminuição da demanda. Com uma demanda 20% menor, os leitos disponíveis são mais que suficientes, diminuindo a taxa de admissões recusadas a 2%. Assim como na maternidade $M_1$, à medida que o tempo de permanência das gestantes diminui, também diminui a necessidade de um número maior de leitos para partos.

Para o segundo, cenário onde as maternidades possuem com uma carga menor, analisado na seção 4, também foram feitas simulações para o dimensionamento do número de leitos operacionais. A demanda nesse cenário foi 25% menor em relação ao primeiro cenário. O número de leitos fixados também foram menores em cerca de 30%. Portanto, para a maternidade $M_1$ o número de leitos operacionais fixado para simulação foi de 33 e 21 para maternidade $M_2$. A Tabela 5.4 apresenta o número de leitos requeridos para diferentes taxas de rejeição na maternidade $M_1$.

Tabela 5.4 – Número de leitos necessários para diferentes taxas de rejeição na Maternidade $M_1$.

| **Número de chegadas diárias ($\lambda$)** | **Tempo médio de permanência (dias)** | **Nº leitos requeridos para rejeição de:** | | | | | |
|---|---|---|---|---|---|---|---|
| | | **25%** | **20%** | **15%** | **10%** | **5%** | **2%** |
| 10.39 | 3,5* | 29 | 32 | 34 | 37 | **41** | 45 |
| | 2,53 (PR) | 21 | 24 | 26 | 28 | **31** | 34 |
| | 1,52 (R) | 13 | 15 | 16 | 18 | **20** | 23 |
| | 1,64 (T) | 14 | 16 | 17 | 19 | **21** | 24 |
| | 1,40 (TP) | 13 | 14 | 15 | 17 | **19** | 21 |
| | 1,54 (L5) | 14 | 15 | 16 | 18 | **20** | 23 |
| | 1,84 (L10) | 16 | 18 | 19 | 21 | **24** | 26 |
| | 1,93 (JSQ) | 17 | 18 | 20 | 22 | **25** | 27 |
| | 1,38 (HBB-LB) | 12 | 14 | 15 | 16 | **19** | 21 |

*Valores reais estimados

Para esse cenário, o número de leitos operacionais fixados para a maternidade $M_1$ não são suficientes para se atingir uma meta de 5% de admissões recusadas. Seriam necessários mais 8 leitos ou uma diminuição de 20% na demanda diária da maternidade. Se o tempo médio de permanência diminuísse cerca de 27% se aproximando do tempo médio observado pelo SUS que foi de 2,3 dias, a quantidade de leitos não necessitaria ser modificada. O tempo médio de permanência está mais relacionado à taxa de parto cirúrgico e de gestações de alto risco, pois a recuperação da parturiente que realizou parto normal é

geralmente mais rápida do que aquela que realizou parto cirúrgico. Para a maternidade $M_2$, o número requerido de leitos é apresentado na Tabela 5.5.

Tabela 5.5 – Número de leitos necessários para diferentes taxas de rejeição na Maternidade $M_2$.

| **Número de chegadas diárias ($\lambda$)** | **Tempo médio de permanência (dias)** | **Nº leitos requeridos para rejeição de:** | | | | | |
|---|---|---|---|---|---|---|---|
| | | **25%** | **20%** | **15%** | **10%** | **5%** | **2%** |
| 6.77 | 3* | 17 | 18 | 20 | 22 | **25** | 28 |
| | 2,08 (PR) | 12 | 13 | 15 | 16 | **18** | 21 |
| | 1,35 (R) | 8 | 9 | 10 | 11 | **13** | 15 |
| | 1,32 (T) | 8 | 9 | 10 | 11 | **13** | 15 |
| | 1,16 (TP) | 7 | 8 | 9 | 10 | **12** | 13 |
| | 1,61 (L5) | 10 | 11 | 12 | 13 | **15** | 17 |
| | 2,42 (L10) | 14 | 15 | 17 | 18 | **21** | 23 |
| | 1,46 (JSQ) | 9 | 10 | 11 | 12 | **14** | 16 |
| | 1,15 (HBB-LB) | 7 | 8 | 9 | 11 | **12** | 13 |

*Valores reais estimados

É possível observar que com uma demanda 25% menor, o número de leitos operacionais requeridos diminui em até 70%. No entanto, na prática, em uma maternidade de uma grande cidade a demanda por serviços de parto é alta e, portanto, há a necessidade de um número maior de leitos. Nota-se também que para atingir a meta de 5% de admissões recusadas é necessário aumentar em 4 o número de leitos operacionais. Como nem sempre é possível aumentar de imediato a capacidade das maternidades, a falta de leitos trará restrições às diversas etapas de realização dos partos levando a uma sobrecarga de gestantes aguardando por leitos, sejam anteriores ou até mesmo após o parto. Uma boa alternativa para suprir a dificuldade em se alterar o número de leitos é adequar a quantidade de partos à capacidade física das maternidades.

A minimização das taxas de admissões recusadas nas maternidades está diretamente ligada a programação dos recursos, ao fluxo de gestantes, ao planejamento da capacidade e aos modelos de alocação de recursos. Através dos resultados obtidos, comprovou-se que, o planejamento da capacidade das maternidades leva a uma grande melhoria no desempenho dessas unidades. Porém, manter as taxas de admissões recusadas baixas acaba se tornando um grande desafio, pois exige que as taxas de ocupação sejam consideravelmente baixas, o que traz outro problema, pois menores taxas de ocupação aumentam a ociosidade de recursos físicos e humanos. Além disso, nas unidades de serviço perinatal as chegadas de gestantes são frequentemente aleatórias com tempos de serviço estocásticos o que tornam a rede um sistema estocástico e o planejamento da capacidade uma questão delicada e complexa (PEHLIVAN, 2014).

Os resultados obtidos pelo modelo de perda de Erlang, usado para descrever o fluxo

de gestantes nas duas maternidades, trouxe uma descrição precisa do número de leitos ocupados e do número de leitos necessários de acordo com a demanda exigida. Mostrou também a forte relação entre a capacidade de uma unidade perinatal, o tempo médio de permanência das gestantes e a probabilidade de uma admissão ser recusada. Em unidades onde o número de leitos é pequeno em comparação a uma elevada demanda por serviços perinatais a probabilidade de admissões recusadas tende a ser alta, comprometendo o gerenciamento de tais unidades. Por outro lado, unidades com uma capacidade operacional grande e uma demanda baixa leva a ociosidade de recursos ocasionando maiores custos e uma baixa eficiência da unidade. No entanto, o modelo permitiu equilibrar o número de leitos operacionais de acordo com demanda exigida facilitando assim uma boa opção de provisão de leitos, e alto acesso ao serviço. Assim, é possível calcular o número ideal de leitos, dado um baixo nível de rejeição de gestantes.

# 6 Conclusão

Nesta dissertação, foi considerado o processo de admissão e escalonamento de gestantes em uma rede perinatal composta por duas maternidades. Foram descritos os principais setores de serviços que compõem cada unidade perinatal e suas respectivas funções e recursos utilizados. Além disso, foi feito um levantamento dos principais desafios e dificuldades enfrentados por essas unidades de saúde no Brasil nas últimas décadas. Assim, diante dos problemas e desafios apontados, foram utilizados diferentes algoritmos de balanceamento de carga na rede perinatal proposta, a fim de encontrar a melhor política de escalonamento de tarefas no sistema que aumente a eficiência da rede. Modelos de filas para um sistema simples de múltiplos processadores heterogêneos foram apresentados, analisados e aplicados a um sistema com duas maternidades. Esses modelos que incorporam estratégias de roteamento de trabalho adaptáveis reduziram o tempo médio de trabalho equilibrando a carga total entre estas duas maternidades e entre os principais setores de cada uma delas.

O algoritmo HBB-LB, baseado no comportamento das abelhas, foi o mais eficiente, já que proporcionou o menor tempo médio de trabalho. Os resultados obtidos confirmaram que, políticas de roteamento e escalonamento que levam em consideração a taxa de chegadas de tarefas e o comprimento de fila do sistema são mais eficientes à medida que a taxa de chegadas aumenta, sendo, portanto, aplicáveis em sistemas de saúde com uma crescente demanda. A média de permanência obtida nas simulações apresentaram valores menores que os dados reais apresentados pelo SUS. Os resultados diminuíram o tempo de permanência das gestantes na rede perinatal, se considerando o tempo médio de ambos os partos e gestações de baixo e alto risco. No melhor cenário obteve-se uma diminuição de cerca de 50% no tempo de permanência. Portanto, em um sistema que enfrenta problemas de alocação de pacientes devido a grande demanda de atendimentos e uma quantidade limitada de recursos, os algoritmos de roteamento podem ser utilizados a fim de melhorar o desempenho de sistemas de saúde.

Foi considerada também a aplicação do modelo de perda de Erlang M/G/$m$/$m$ para descrever o fluxo de gestantes que chegam às maternidades. Utilizou-se os dados de duas maternidades que descrevem o processo de chegadas e informam o tempo de permanência e a capacidade de internação em leitos para partos. O modelo de Erlang foi então utilizado para o dimensionamento da capacidade das duas maternidades analisadas. Foi possível estimar o número requerido de leitos variando-se as taxas de admissões de gestantes recusadas, a demanda exigida e o tempo de permanência das gestantes em cada uma das duas maternidades. Provou-se que, para os tempos de permanência obtidos pelos algoritmos de balanceamento de carga, consegue-se alcançar uma maior eficiência na

utilização dos leitos operacionais, garantindo baixas taxas de rejeição e uma capacidade operacional menor, o que contribui para diminuição de custos.

Como trabalho futuro, pretende-se inserir mais variáveis ao modelo, já que nesta dissertação não foram considerados as taxas de mortalidade neonatal e materna, além das taxas de rejeição na rede. Pretende-se encontrar melhores políticas de escalonamento de gestantes e estudar o planejamento da localização geográfica de forma a diminuir as taxas de rejeição.

# Referências


Abbass, H. A. Mbo: marriage in honey bees optimization-a haplometrosis polygynous swarming approach. In: *Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No.01TH8546)*. [S.l.: s.n.], 2001. v. 1, p. 207–214 vol. 1. Citado na página 41.

ABRAHAM, A.; GUO, H.; LIU, H. Swarm intelligence: foundations, perspectives and applications. In: *Swarm intelligent systems*. [S.l.]: Springer, 2006. p. 3–25. Citado na página 39.

ALMEIDA, W. d. S. d.; SZWARCWALD, C. L. Mortalidade infantil e acesso geográfico ao parto nos municípios brasileiros. *Revista de Saúde Pública*, SciELO Public Health, v. 46, p. 68–76, 2012. Citado na página 19.

ALVES, M. T. S. S. d. et al. Atenção ao aborto no sistema único de saúde no nordeste brasileiro: a estrutura dos serviços. *Revista Brasileira de Saúde Materno Infantil*, SciELO Brasil, v. 14, n. 3, p. 229–239, 2014. Citado na página 19.

ARMONY, M. et al. On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic systems*, INFORMS, v. 5, n. 1, p. 146–194, 2015. Citado na página 77.

AUGUSTO, V.; MURGIER, M.; VIALLON, A. A modelling and simulation framework for intelligent control of emergency units in the case of major crisis. In: IEEE. *2018 Winter Simulation Conference (WSC)*. [S.l.], 2018. p. 2495–2506. Citado na página 21.

BABU, L. D. D.; KRISHNA, P. V. Honey bee behavior inspired load balancing of tasks in cloud computing environments. In: *Applied Soft Computing Journal*. [S.l.]: Elsevier, 2013. p. 2292–2303. Citado na página 42.

BANDYOPADHYAY, S. *Dissemination of Information in Optical Networks:: From Technology to Algorithms*. [S.l.]: Springer Science & Business Media, 2007. Citado na página 45.

BELCIUG, S.; GORUNESCU, F. Improving hospital bed occupancy and resource utilization through queuing modeling and evolutionary computation. *Journal of biomedical informatics*, Elsevier, v. 53, p. 261–269, 2015. Citado na página 78.

BERNARDINO, A. M. et al. Efficient load balancing using the bees algorithm. In: SPRINGER. *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. [S.l.], 2011. p. 469–479. Citado na página 42.

BITTENCOURT, S. D. d. A. et al. Estrutura das maternidades: aspectos relevantes para a qualidade da atenção ao parto e nascimento. *Cadernos de Saúde Pública*, SciELO Public Health, v. 30, p. S208–S219, 2014. Citado na página 19.

BOLCH, G. et al. *Queueing networks and Markov chains: modeling and performance evaluation with computer science applications*. [S.l.]: John Wiley & Sons, 2006. Citado 2 vezes nas páginas 52 e 55.

BONABEAU, E. et al. *Swarm intelligence: from natural to artificial systems*. [S.l.]: Oxford university press, 1999. Citado 2 vezes nas páginas 38 e 39.

BRANDEAU, M. L.; SAINFORT, F.; PIERSKALLA, W. P. *Operations research and health care: a handbook of methods and applications*. [S.l.]: Springer Science & Business Media, 2004. v. 70. Citado 2 vezes nas páginas 23 e 45.

BRUIN, A. M. D. et al. Modeling the emergency cardiac in-patient flow: an application of queuing theory. *Health Care Management Science*, Springer, v. 10, n. 2, p. 125–137, 2007. Citado na página 77.

BRUIN, A. M. de et al. Dimensioning hospital wards using the erlang loss model. *Annals of Operations Research*, Springer, v. 178, n. 1, p. 23–43, 2010. Citado 4 vezes nas páginas 77, 78, 80 e 84.

CABRAL, F. B. The slow server problem for uninformed customers. *Queueing systems*, Springer, v. 50, n. 4, p. 353–370, 2005. Citado na página 37.

CAMPOS, T. P.; CARVALHO, M. S. Assistência ao parto no município do rio de janeiro: perfil das maternidades e o acesso da clientela. *Cadernos de Saúde Pública*, SciELO Brasil, v. 16, n. 2, p. 411–420, 2000. Citado na página 19.

CASTRO, L. N. D. *Fundamentals of natural computing: basic concepts, algorithms, and applications*. [S.l.]: CRC Press, 2006. Citado na página 41.

CASTRO, L. N. D.; ZUBEN, F. J. V. Artificial immune systems: Part i–basic theory and applications. *Universidade Estadual de Campinas, Dezembro de, Tech. Rep*, v. 210, n. 1, 1999. Citado na página 39.

CHOW, Y.-C.; KOHLER, W. H. Dynamic load balancing in homogeneous two-processor distributed systems. In: *Proceedings of the International Symposium on Computer Performance, Modeling, Measurement and Evaluation*. [S.l.: s.n.], 1977. p. 39–52. Citado na página 35.

CHOW, Y.-C.; KOHLER, W. H. Models for dynamic load balancing in a heterogeneous multiple processor system. *IEEE Transactions on Computers*, IEEE, v. 100, n. 5, p. 354–361, 1979. Citado 7 vezes nas páginas 9, 36, 47, 48, 49, 51 e 52.

CLERC, M.; KENNEDY, J. The particle swarm-explosion, stability, and convergence in a multidimensional complex space. *IEEE transactions on Evolutionary Computation*, IEEE, v. 6, n. 1, p. 58–73, 2002. Citado na página 39.

COSTA, G. D. d. et al. Avaliação do cuidado à saúde da gestante no contexto do programa saúde da família. *Ciência & saúde coletiva*, SciELO Public Health, v. 14, p. 1347–1357, 2009. Citado na página 19.

CRUISE, J. et al. Stability of jsq in queues with general server-job class compatibilities. *Queueing Systems: Theory and Applications*, Springer, p. 1–9, 2020. Citado na página 35.

DORIGO, M. Optimization, learning and natural algorithms. *PhD Thesis, Politecnico di Milano*, 1992. Citado na página 39.

DORIGO, M.; MANIEZZO, V.; COLORNI, A. Ant system: optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, IEEE, v. 26, n. 1, p. 29–41, 1996. Citado na página 39.

EAGER, D. L.; LAZOWSKA, E. D.; ZAHORJAN, J. Adaptive load sharing in homogeneous distributed systems. *IEEE transactions on software engineering*, IEEE, n. 5, p. 662–675, 1986. Citado na página 36.

EFFATPARVAR, M.; GARSHASBI, M. A genetic algorithm for static load balancing in parallel heterogeneous systems. *Procedia-Social and Behavioral Sciences*, Elsevier, v. 129, p. 358–364, 2014. Citado na página 37.

ESSEN, J. T. van; HOUDENHOVEN, M. van; HURINK, J. L. Clustering clinical departments for wards to achieve a prespecified blocking probability. *OR spectrum*, Springer, v. 37, n. 1, p. 243–271, 2015. Citado 2 vezes nas páginas 78 e 81.

GAMBARDELLA, L. M.; TAILLARD, É.; AGAZZI, G. Macs-vrptw: A multiple colony system for vehicle routing problems with time windows. In: CITESEER. *New ideas in optimization*. [S.l.], 1999. Citado na página 39.

GORUNESCU, F.; MCCLEAN, S. I.; MILLARD, P. H. A queueing model for bed-occupancy management and planning of hospitals. *Journal of the Operational Research Society*, Springer, v. 53, n. 1, p. 19–24, 2002. Citado na página 78.

GREEN, L. V. How many hospital beds? *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, SAGE Publications Sage CA: Los Angeles, CA, v. 39, n. 4, p. 400–412, 2002. Citado na página 78.

GREEN, L. V. Capacity planning and management in hospitals. In: *Operations research and health care*. [S.l.]: Springer, 2005. p. 15–41. Citado na página 20.

GREEN, L. V. Using operations research to reduce delays for healthcare. In: *State-of-the-Art Decision-Making Tools in the Information-Intensive Age*. [S.l.]: INFORMS, 2008. p. 1–16. Citado na página 20.

GREEN, L. V.; NGUYEN, V. Strategies for cutting hospital beds: the impact on patient service. *Health services research*, Health Research & Educational Trust, v. 36, n. 2, p. 421, 2001. Citado na página 78.

GREEN, L. V. et al. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine*, Wiley Online Library, v. 13, n. 1, p. 61–68, 2006. Citado na página 77.

GROSU, D.; CHRONOPOULOS, A. T. Noncooperative load balancing in distributed systems. *Journal of parallel and distributed computing*, Elsevier, v. 65, n. 9, p. 1022–1034, 2005. Citado na página 35.

GROSU, D.; CHRONOPOULOS, A. T.; LEUNG, M.-Y. Load balancing in distributed systems: An approach using cooperative games. In: IEEE. *Proceedings 16th International Parallel and Distributed Processing Symposium*. [S.l.], 2002. p. 10–pp. Citado na página 37.

GULLIFORD, M. et al. What does' access to health care'mean? *Journal of health services research & policy*, SAGE Publications Sage UK: London, England, v. 7, n. 3, p. 186–188, 2002. Citado na página 27.

Gupta, E.; Deshpande, V. A technique based on ant colony optimization for load balancing in cloud data center. In: *International Conference on Information Technology*. [S.l.: s.n.], 2014. p. 12–17. Citado na página 39.

GUPTA, V. et al. Analysis of join-the-shortest-queue routing for web server farms. *Performance Evaluation*, Elsevier, v. 64, n. 9-12, p. 1062–1081, 2007. Citado na página 37.

HADDAD, O. B.; AFSHAR, A. Mbo (marriage bees optimization) algorithm, a new heuristic approach in hydrosystems design and operation. 2004. Citado na página 41.

HADDAD, O. B.; AFSHAR, A.; MARINO, M. A. Honey-bees mating optimization (hbmo) algorithm: a new heuristic approach for water resources optimization. *water resources management*, Springer, v. 20, n. 5, p. 661–680, 2006. Citado na página 41.

HARPER, P. R.; SHAHANI, A. Modelling for the planning and management of bed capacities in hospitals. *Journal of the Operational research Society*, Taylor & Francis, v. 53, n. 1, p. 11–18, 2002. Citado na página 78.

HIGHTOWER, R. R.; FORREST, S.; PERELSON, A. S. The evolution of emergent organization in immune system gene libraries. In: CITESEER. *ICGA*. [S.l.], 1995. p. 344–350. Citado na página 39.

IBGE. *Estatísticas do Registro Civil*. 2018. <https://www.ibge.gov.br/estatisticas/todos-os-produtos-estatisticas/26177-pareamento-de-dados.html?edicao=26191&t=downloads>. Acessado: 14-02-2020. Citado na página 28.

ILIADIS, I.; LIEN, L.-C. Resequencing delay for a queueing system with two heterogeneous servers under a threshold-type scheduling. *IEEE Transactions on Communications*, IEEE, v. 36, n. 6, p. 692–702, 1988. Citado na página 57.

JONES, R. A simple guide to a complex problem—maternity bed occupancy. *British Journal of Midwifery*, MA Healthcare London, v. 20, n. 5, p. 351–357, 2012. Citado na página 77.

JUNG, S. H. Queen-bee evolution for genetic algorithms. *Electronics letters*, IET, v. 39, n. 6, p. 575–576, 2003. Citado na página 41.

KARABOGA, D. An idea based on honey bee swarm for numerical optimization. In: *Tech. Report TR06*. Erciyes University, Engineering Faculty, Computer Engineering Department: [s.n.], 2005. Citado 2 vezes nas páginas 39 e 41.

KESKINTURK, T.; YILDIRIM, M. B.; BARUT, M. An ant colony optimization algorithm for load balancing in parallel machines with sequence-dependent setup times. *Computers & Operations Research*, Elsevier, v. 39, n. 6, p. 1225–1235, 2012. Citado na página 39.

KLEINROCK, L. *Queueing systems. Volume I: theory*. [S.l.]: wiley New York, 1975. Citado 3 vezes nas páginas 57, 59 e 82.

KOKANGUL, A. A combination of deterministic and stochastic approaches to optimize bed capacity in a hospital unit. *Computer methods and programs in biomedicine*, Elsevier, v. 90, n. 1, p. 56–65, 2008. Citado na página 78.

KRISHNAMOORTHI, B. On poisson queue with two heterogeneous servers. *Operations Research*, INFORMS, v. 11, n. 3, p. 321–330, 1963. Citado na página 57.

LAFFEL, G.; BLUMENTHAL, D. The case for using industrial quality management science in health care organizations. *Jama*, American Medical Association, v. 262, n. 20, p. 2869–2873, 1989. Citado na página 77.

LARSEN, R. L.; AGRAWALA, A. K. Control of a heterogeneous two-server exponential queueing system. *IEEE Transactions on Software Engineering*, IEEE, n. 4, p. 522–526, 1983. Citado 2 vezes nas páginas 37 e 56.

LEAL, M. do C. et al. Atenção ao pré-natal e parto em mulheres usuárias do sistema público de saúde residentes na amazônia legal e no nordeste, Brasil 2010. *Revista Brasileira de Saúde Materno Infantil*, v. 15, n. 1, p. 91–104, 2015. Citado na página 19.

LI, K. et al. Cloud task scheduling based on load balancing ant colony optimization. In: IEEE. *2011 sixth annual ChinaGrid conference*. [S.l.], 2011. p. 3–9. Citado na página 39.

LI, X. et al. An integrated queuing and multi-objective bed allocation model with application to a hospital in china. *Journal of the Operational Research Society*, Springer, v. 60, n. 3, p. 330–338, 2009. Citado na página 78.

LIN, H.-C.; RAGHAVENDRA, C. S. A dynamic load-balancing policy with a central job dispatcher (lbc). *IEEE Transactions on Software Engineering*, IEEE, n. 2, p. 148–158, 1992. Citado na página 35.

LIN, H.-C.; RAGHAVENDRA, C. S. A state-aggregation method for analyzing dynamic load-balancing policies. In: IEEE. *[1993] Proceedings. The 13th International Conference on Distributed Computing Systems*. [S.l.], 1993. p. 482–489. Citado na página 36.

LIN, H.-C.; RAGHAVENDRA, C. S. An approximate analysis of the join the shortest queue (jsq) policy. *IEEE Transactions on Parallel and Distributed Systems*, IEEE, v. 7, n. 3, p. 301–307, 1996. Citado 2 vezes nas páginas 37 e 60.

LIN, W.; KUMAR, P. Optimal control of a queueing system with two heterogeneous servers. *IEEE Transactions on Automatic control*, IEEE, v. 29, n. 8, p. 696–703, 1984. Citado 5 vezes nas páginas 9, 37, 56, 57 e 58.

LIVNY, M.; MELMAN, M. Load balancing in homogeneous broadcast distributed systems. In: *Proceedings of the Computer Network Performance Symposium*. [S.l.: s.n.], 1982. p. 47–55. Citado na página 36.

LU, X.; ZHOU, Y. A novel global convergence algorithm: bee collecting pollen algorithm. In: SPRINGER. *International Conference on Intelligent Computing*. [S.l.], 2008. p. 518–525. Citado na página 42.

LUCIC, P.; TEODOROVIć, D. Bee system: Modeling combinatorial optimization transportation engineering problems by swarm intelligence. *Preprints of the TRISTAN IV Triennial Symposium on Transportation Analysis*, p. 441–445, 01 2001. Citado na página 41.

MANDELBAUM, A.; MOMČILOVIĆ, P.; TSEYTLIN, Y. On fair routing from emergency departments to hospital wards: Qed queues with heterogeneous servers. *Management Science*, INFORMS, v. 58, n. 7, p. 1273–1291, 2012. Citado na página 77.

MESTRE, A. M.; OLIVEIRA, M. D.; BARBOSA-PÓVOA, A. Organizing hospitals into networks: a hierarchical and multiservice model to define location, supply and referrals in planned hospital systems. *OR spectrum*, Springer, v. 34, n. 2, p. 319–348, 2012. Citado na página 20.

MILLONAS, M. M. Swarms, phase transitions, and collective intelligence. In: ADDISON-WESLEY. *Langton C (eds), Artificial life III*. [S.l.], 1994. p. 417–445. Citado na página 40.

Ministério da Saúde. *Boletim Epidemiológico*. 2019. <https://portalarquivos2.saude.gov.br/images/pdf/2019/setembro/25/boletim-especial-21ago19-web.pdf>. Acessado em 10 de Novembro 2020. Citado na página 30.

MIRCHANDANEY, R.; TOWSLEY, D.; STANKOVIC, J. A. Analysis of the effects of delays on load sharing. *IEEE transactions on computers*, IEEE, v. 38, n. 11, p. 1513–1525, 1989. Citado na página 36.

MUKHOPADHYAY, A.; MAZUMDAR, R. R. Analysis of randomized join-the-shortest-queue (jsq) schemes in large heterogeneous processor-sharing systems. *IEEE Transactions on Control of Network Systems*, IEEE, v. 3, n. 2, p. 116–126, 2015. Citado na página 60.

NAKRANI, S.; TOVEY, C. On honey bees and dynamic allocation in an internet server colony. In: CITESEER. *Proceedings of 2nd International Workshop on the Mathematics and Algorithms of Social Insects*. [S.l.], 2003. p. 1–8. Citado na página 41.

NELSON, R. D.; PHILIPS, T. K. An approximation to the response time for shortest queue routing. *ACM SIGMETRICS Performance Evaluation Review*, ACM New York, NY, USA, v. 17, n. 1, p. 181–189, 1989. Citado na página 37.

NI, L. M.; HWANG, K. Optimal load balancing in a multiple processor system with many job classes. *IEEE Transactions on Software Engineering*, IEEE, n. 5, p. 491–496, 1985. Citado 2 vezes nas páginas 36 e 57.

NISHANT, K. et al. Load balancing of nodes in cloud using ant colony optimization. In: IEEE. *2012 UKSim 14th international conference on computer modelling and simulation*. [S.l.], 2012. p. 3–8. Citado na página 39.

OMS. *Declaração da OMS sobre Taxas de Cesáreas*. 2015. <https://apps.who.int/iris/bitstream/handle/10665/161442/WHO_RHR_15.02_por.pdf?sequence=3>. Acessado em 07 de Janeiro 2021. Citado na página 33.

PARPINELLI, R. S.; LOPES, H. S. New inspirations in swarm intelligence: a survey. *International Journal of Bio-Inspired Computation*, Inderscience Publishers, v. 3, n. 1, p. 1–16, 2011. Citado na página 43.

PARSOPOULOS, K. E.; VRAHATIS, M. N. On the computation of all global minimizers through particle swarm optimization. *IEEE Transactions on evolutionary computation*, IEEE, v. 8, n. 3, p. 211–224, 2004. Citado na página 39.

PEHLIVAN, C. *Design and flow control of stochastic health care networks without waiting rooms: A perinatal application*. Tese (Doutorado) — Ecole Nationale Supérieure des Mines de Saint-Etienne, 2014. Citado 4 vezes nas páginas 20, 29, 77 e 87.

PEHLIVAN, C.; AUGUSTO, V.; XIE, X. Admission control in a pure loss healthcare network: Mdp and des aproach. *Proceedings of the 2013 Winter Simulation Conference*, v. 32, p. 5–44, 2013. Citado 2 vezes nas páginas 20 e 78.

PEHLIVAN, C.; AUGUSTO, V.; XIE, X. Dynamic capacity planning and location of hierarchical service networks under service level constraints. *IEEE Transactions on Automation Science and Engineering*, IEEE, v. 11, n. 3, p. 863–880, 2014. Citado na página 77.

PEHLIVAN, C. et al. Multi-period capacity planning for maternity facilities in a perinatal network: A queuing and optimization approach. In: IEEE. *2012 IEEE International Conference on Automation Science and Engineering (CASE)*. [S.l.], 2012. p. 137–142. Citado na página 77.

PEREIRA, R. M. et al. Novas práticas de atenção ao parto e os desafios para a humanização da assistência nas regiões sul e sudeste do Brasil. *Ciência & Saúde Coletiva*, SciELO Public Health, v. 23, p. 3517–3524, 2018. Citado na página 19.

PHAM, D. T. et al. The bees algorithm—a novel tool for complex optimisation problems. In: *Intelligent production machines and systems*. [S.l.]: Elsevier, 2006. p. 454–459. Citado na página 42.

PRACTICE, A. C. on O. et al. *Guidelines for perinatal care*. [S.l.]: Am Acad Pediatrics, 2012. Citado 2 vezes nas páginas 32 e 33.

PRODEL, M.; AUGUSTO, V.; XIE, X. Hospitalization admission control of emergency patients using markovian decision processes and discrete event simulation. In: IEEE. *Proceedings of the Winter Simulation Conference 2014*. [S.l.], 2014. p. 1433–1444. Citado na página 21.

RAMAKRISHNAN, M.; SIER, D.; TAYLOR, P. A two-time-scale model for hospital patient flow. *IMA Journal of Management Mathematics*, Oxford University Press, v. 16, n. 3, p. 197–215, 2005. Citado na página 21.

RêGO, M. G. d. S. et al. Óbitos perinatais evitáveis por intervenções do sistema Único de saúde do Brasil. *Revista Gaúcha de Enfermagem*, scielo, v. 39, n. e2017-0084, 2018. ISSN 1983-1447. Citado na página 20.

SAúDE, M. da. *Ministério da Saúde. TabNet Win32 3.0: Internações hospitalares do SUS - por local de internação - SUS*. 2012. <http://tabnet.datasus.gov.br/cgi/tabcgi.exe?sih/cnv/sxuf.def>. Acessado: 14-02-2020. Citado 3 vezes nas páginas 31, 69 e 84.

SAúDE, M. da. *Ministério da Saúde. Cadastro Nacional de Estabelecimentos de Saúde CNES*. 2019. <http://tabnet.datasus.gov.br/cgi/deftohtm.exe?cnes/cnv/leiintbr.def>. Acessado: 14-02-2020. Citado na página 27.

SELEN, J. et al. Steady-state analysis of shortest expected delay routing. *Queueing Systems*, Springer, v. 84, n. 3-4, p. 309–354, 2016. Citado na página 35.

SERAPIÃO, A. B. d. S. Fundamentos de otimização por inteligência de enxames: uma visão geral. *Sba: Controle & Automação Sociedade Brasileira de Automatica*, SciELO Brasil, v. 20, n. 3, p. 271–304, 2009. Citado 3 vezes nas páginas 38, 40 e 43.

SILVA, A. L. A. d. et al. A qualidade do atendimento ao parto na rede pública hospitalar em uma capital brasileira: a satisfação das gestantes. *Cadernos de Saúde Pública*, SciELO Public Health, v. 33, p. e00175116, 2017. Citado na página 19.

SMITH, D. J. et al. Using lazy evaluation to simulate realistic-size repertoires in models of the immune system. *Bulletin of Mathematical Biology*, Springer, v. 60, n. 4, p. 647–658, 1998. Citado na página 39.

STOCKBRIDGE, R. H. A martingale approach to the slow server problem. *Journal of applied probability*, Cambridge University Press, v. 28, n. 2, p. 480–486, 1991. Citado na página 37.

SUBRATA, R.; ZOMAYA, A. Y.; LANDFELDT, B. Game-theoretic approach for load balancing in computational grids. *IEEE Transactions on Parallel and Distributed Systems*, IEEE, v. 19, n. 1, p. 66–76, 2007. Citado na página 37.

TSEYTLIN, Y. *Queueing systems with heterogeneous servers: On fair routing of patients in emergency departments*. Tese (Doutorado) — Technion-Israel Institute of Technology, Faculty of Industrial and . . . , 2009. Citado na página 77.

TSEYTLIN, Y.; ZVIRAN, A. Simulation of patients routing from an emergency department to internal wards in rambam hospital. *OR Graduate Project, IE&M, Technion*, 2008. Citado 2 vezes nas páginas 19 e 20.

UNGLERT, C. V. d. S. O enfoque da acessibilidade no planejamento da localização e dimensão de serviços de saúde. *Revista de Saúde Pública*, SciELO Brasil, v. 24, n. 6, p. 445–452, 1990. Citado na página 19.

VÉRICOURT, F. d.; JENNINGS, O. B. Nurse staffing in medical units: A queueing perspective. *Operations Research*, INFORMS, v. 59, n. 6, p. 1320–1331, 2011. Citado na página 77.

VESTERSTRØM, J.; RIGET, J. Particle swarms: Extensions for improved local, multi-modal, and dynamic search in numerical optimization. *Master's Thesis, May*, 2002. Citado na página 39.

VOGEL, J. P. et al. Use of the robson classification to assess caesarean section trends in 21 countries: a secondary analysis of two who multicountry surveys. *The Lancet Global Health*, Elsevier, v. 3, n. 5, p. e260–e270, 2015. Citado na página 33.

WEDDE, H. F.; FAROOQ, M.; ZHANG, Y. Beehive: An efficient fault-tolerant routing algorithm inspired by honey bee behavior. In: SPRINGER. *International Workshop on Ant Colony Optimization and Swarm Intelligence*. [S.l.], 2004. p. 83–94. Citado na página 41.

WOLFF, R. W. *Stochastic modeling and the theory of queues*. [S.l.]: Pearson College Division, 1989. Citado na página 80.

YANG, X.-S. Engineering optimizations via nature-inspired virtual bee algorithms. In: SPRINGER. *International Work-Conference on the Interplay Between Natural and Artificial Computation*. [S.l.], 2005. p. 317–323. Citado na página 41.

YANKOVIC, N.; GREEN, L. V. Identifying good nursing levels: A queuing approach. *Operations research*, INFORMS, v. 59, n. 4, p. 942–955, 2011. Citado na página 77.

ZAIED, I. *The Offered Load in Fork-Join Networks: Calculations and Applications to Service Engineering of Emergency Department. M. Sc. Research Proposal*. Tese (Doutorado) — Technion-Israel Institute of Technology, 2010. Citado na página 77.

ZOMAYA, A. Y.; TEH, Y.-H. Observations on using genetic algorithms for dynamic load-balancing. *IEEE transactions on parallel and distributed systems*, IEEE, v. 12, n. 9, p. 899–911, 2001. Citado na página 37.