UNIVERSIDADE ESTADUAL PAULISTA
JÚLIO DE MESQUITA FILHO
CAMPUS DE BAURU – FACULDADE DE ENGENHARIA
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO

CAIO VITOR BEOJONE

AVALIAÇÃO DO DESEMPENHO E CENÁRIOS ALTERNATIVOS EM UM SAMU UTILIZANDO O MODELO HIPERCUBO ESTACIONÁRIO E NÃO-ESTACIONÁRIO

BAURU
2017

UNIVERSIDADE ESTADUAL PAULISTA
JÚLIO DE MESQUITA FILHO
CAMPUS DE BAURU – FACULDADE DE ENGENHARIA
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO

CAIO VITOR BEOJONE

AVALIAÇÃO DO DESEMPENHO E CENÁRIOS ALTERNATIVOS EM UM SAMU UTILIZANDO O MODELO HIPERCUBO ESTACIONÁRIO E NÃO-ESTACIONÁRIO

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Estadual Paulista, como parte dos requisitos para obtenção do título de Mestre em Engenharia de Produção.
Orientadora: Profa. Dra. Regiane Máximo de Souza.

BAURU
2017

Beojone, Caio Vítor.
Avaliação do desempenho e cenários alternativos em um SAMU utilizando o modelo hipercubo estacionário e não-estacionário / Caio Vítor Beojone, 2017
173 f. : il.

Orientador: Regiane Máximo de Souza

Dissertação (Mestrado)–Universidade Estadual Paulista. Faculdade de Engenharia, Bauru, 2017

1. Teoria das filas. 2. Pesquisa operacional em saúde. 3. Serviço Médico Emergencial. 4. Sistemas não-estacionários. 5. Reserva de capacidade. I. Universidade Estadual Paulista. Faculdade de Engenharia. II. Título.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Bauru

ATA DA DEFESA PÚBLICA DA DISSERTAÇÃO DE MESTRADO DE CAIO VÍTOR BEOJONE, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO, DA FACULDADE DE ENGENHARIA - CÂMPUS DE BAURU.

Aos 09 dias do mês de outubro do ano de 2017, às 14:30 horas, no(a) Anfiteatro da Seção Técnica de Pós-graduação da FEB, reuniu-se a Comissão Examinadora da Defesa Pública, composta pelos seguintes membros: Profa. Dra. REGIANE MÁXIMO DE SOUZA - Orientador(a) do(a) Departamento de Engenharia de Produção / Faculdade de Engenharia de Bauru - UNESP, Prof. Dr. ENZO BARBERIO MARIANO do(a) Departamento de Engenharia de Produção / Faculdade de Engenharia de Bauru, Prof. Dr. REINALDO MORABITO NETO do(a) Departamento de Engenharia de Produção / Universidade Federal de São Carlos -UFSCar, sob a presidência do primeiro, a fim de proceder a arguição pública da DISSERTAÇÃO DE MESTRADO de CAIO VÍTOR BEOJONE, intitulada **AVALIAÇÃO DO DESEMPENHO E CENÁRIOS ALTERNATIVOS EM UM SAMU UTILIZANDO O MODELO HIPERCUBO ESTACIONÁRIO E NÃO ESTACIONÁRIO**. Após a exposição, o discente foi arguido oralmente pelos membros da Comissão Examinadora, tendo recebido o conceito final:____ Aprovado ____. Nada mais havendo, foi lavrada a presente ata, que após lida e aprovada, foi assinada pelos membros da Comissão Examinadora.

Profa. Dra. REGIANE MÁXIMO DE SOUZA

Prof. Dr. ENZO BARBERIO MARIANO

Prof. Dr. REINALDO MORABITO NETO

RESUMO

Vários Sistemas de Atendimento Emergenciais (SAE's) sofrem com as variações diárias da demanda e da disponibilidade das ambulâncias. Nesses sistemas pode haver flutuação do desempenho ao longo do dia devido, por exemplo, a mudança no número de servidores e nas taxas de chegada, levando à necessidade de considerar explicitamente tais variações em uma extensão ao modelo hipercubo ainda não explorada na literatura. Como ocorre em alguns SAE's, as ambulâncias melhor equipadas são reservadas para o atendimento exclusivo de chamados com risco de vida. Dessa maneira, a política de despacho pode ser diferenciada com a finalidade de reservar totalmente o atendimento de alguns servidores para certas gravidades de ocorrências. Além disso, somam-se à natureza aleatória desses sistemas, como por exemplo, as incertezas da disponibilidade das ambulâncias, a chegada de um novo chamado e sua localização. Nesse contexto, os objetivos do presente estudo são: (i) estender o modelo hipercubo de filas para reserva total de capacidade, dependendo do tipo do chamado; (ii) estender o modelo hipercubo de filas para torná-lo mais eficiente computacionalmente, sem haver perda de precisão durante a modelagem e resolução; e (iii) propor uma abordagem baseada no modelo hipercubo não-estacionário para organização do trabalho das ambulâncias em qualquer momento do dia. Para verificar a viabilidade e a aplicabilidade dessas abordagens, é realizado um estudo de caso no SAMU da cidade de Bauru (SAMU-Bauru) que, além de reservar suas ambulâncias avançadas para ocorrências mais graves, é afetado pelas variações diárias na demanda e disponibilidade das ambulâncias. Além da configuração original do SAMU-Bauru, estudada em duas etapas, foram analisados um total de quatro cenários alternativos que consideram questões importantes: o impacto do aumento na demanda do período mais congestionado; a mitigação desse impacto incluindo uma nova ambulância; a alteração do horário das pausas diárias; e o impacto de aumentos na demanda em horários específicos do dia. Foram calculadas importantes medidas de desempenho para cada cenário como a carga de trabalho, tempos médios de espera e tempos médios de resposta. Os resultados mostram que as extensões realizadas no modelo hipercubo são capazes de analisar satisfatoriamente sistemas como o SAMU-Bauru, além de possibilitar a criação e mensuração de propostas de melhorias nos níveis táticos e operacionais.

Palavras-chave: Teoria das Filas; Pesquisa Operacional em saúde; Serviço Médico Emergencial; Sistemas não-estacionários; Reserva de capacidade.

ABSTRACT

Many Emergency Service Systems face daily variations on demand and ambulance availability. These systems may suffer, for example, performance fluctuations throughout the day, changes on the number of servers and on arrival rates, leading to the need to explicitly consider such variations in a hypercube model extension not yet explored in the literature. As occurs in some SAMU's, which reserve their best equipped ambulances to exclusively serve life-threating requests. Therefore, the dispatch policy can be differentiated in order to completely reserve the service of some ambulances to more severe requests. These problems add up to the random nature of these systems with uncertainties upon ambulance availability or the arrival of a new request and its location. Thus, this study aims to: (i) extend the hypercube queueing model to be able to capture the complete capacity reservation of advanced ambulances, depending on the request classification; (ii) extend the hypercube model in order to make it more computationally efficient, without losing any information during modeling and resolution. (iii) propose an approach based on nonstationary hypercube queueing model to organize the operation of ambulances at any time of the day. To verify the feasibility of these approaches, a case study is carried out on the SAMU from Bauru city (SAMU-Bauru), which, in addition to the advanced ambulance reservation for life-threating requests, is affected by daily variations in demand and ambulance availability. In addition to the original configuration of SAMU-Bauru, studied on a two-step approach, we studied a total of four alternative scenarios that exploited important matters as: the impact of average demand increase on the congestion peak; mitigation of this impact by including a new ambulance; changing the schedule of daily breaks; and the impact of increases in the demand at specific hours of the day. We calculated important performance measures for each scenario, such as workload, mean waiting times and mean response times. Results show that the proposed extensions to the hypercube model are capable of satisfactorily analyze systems such as SAMU-Bauru, besides making it possible to create and to measure improvements proposals in tactical and operational levels.

Keywords: Queueing Theory; Operational Research in Health Services; Emergency Medical Systems; Nonstationary systems; Capacity Reservation.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

PO	Pesquisa Operacional
SAMU	Sistema de Atendimento Móvel de Urgência
SAE	Sistema de Atendimento Emergencial
UTI	Unidade de Terapia Intensiva
VSA	Veículo de Suporte Avançado
VSB	Veículo de Suporte Básico

SUMÁRIO

1 INTRODUÇÃO

Serviços de saúde bem preparados ajudam na diminuição de perdas de vidas e possíveis sequelas dos pacientes. Melhorar a qualidade e a segurança desses sistemas requer um esforço de planejamento e atividade contínuo. A medição do desempenho desses sistemas ajuda a verificar e controlar a qualidade do serviço como o nível de serviço fornecido aos usuários, carga de trabalho, entre outros (PRONOVOST *et al.*, 2009). A gestão dos sistemas de saúde está se tornando mais cara, devido ao surgimento das novas tecnologias e a tendências demográficas, como o envelhecimento da população (BRAILSFORD; VISSERS, 2011). Do ponto de vista de operações, melhorar a eficiência desses sistemas significa diminuir os tempos de espera dos pacientes (PATRICK *et al.*, 2008; LIU; D'AUNNO, 2012). Como forma de melhorar a eficiência, gestores estão reconsiderando o *design* dos serviços de saúde (HULSHOF *et al.*, 2012).

Serviços de Atendimento Emergenciais (SAE's) são críticos na sociedade moderna, provendo assistência aos usuários para incidentes, garantindo saúde e segurança pública (GEROLIMINIS *et al.*, 2011). Por conta disso, seu objetivo primário é prover resposta imediata para chamados de emergência, sejam eles policiais, bombeiros, ou ambulâncias (GALVÃO; MORABITO, 2008). Melhorar a resposta dos SAE's envolve decisões táticas como a localização de suas bases e ambulâncias (RAJAGOPALAN *et al.*, 2008). Tomar tais decisões a partir da experimentação é difícil e caro, já que erros são potencialmente fatais para os usuários (DAVOUDPOUR *et al.*, 2014). Por este motivo, a crescente disponibilidade de informações geográficas (por exemplo, da demanda, da situação dos servidores, etc.) e poder computacional criou condições ideais para se realizar testes computacionais e garantir que as decisões tomadas diminuirão o tempo de resposta aos usuários (MAXWELL *et al.*, 2010).

No Brasil, um dos modelos de atendimentos dos Serviços Médicos Emergenciais em sistemas urbanos é adotado pelo governo federal e é conhecido por SAMU (Serviço de Atendimento Móvel de Urgência). O SAMU brasileiro surgiu a partir de um acordo bilateral assinado por Brasil e França (LOPES; FERNANDES, 1999). Na França, este sistema também é chamado de SAMU (*Service d'Aide Médicale d'Urgence*, em francês) e já opera a mais de 30 anos (TAKEDA *et al.*, 2007). No SAMU, o serviço avançado é realizado primariamente por médicos; ao contrário do modelo americano, em que o serviço é prestado primariamente por paramédicos, profissão não presente no Brasil (SOUZA, 2010). A especialidade para médicos que trabalham com emergências ainda é muito pouco difundida no Brasil e apenas em 2016, através da Resolução CFM nº2.149/2016, passou a ser reconhecida (CFM, 2016).

O SAMU opera 24 horas por dia e seu serviço tem por finalidade prestar socorro emergencial e garantir a qualidade do atendimento. O serviço pode englobar apenas uma cidade ou atender chamados de forma regionalizada. O chamado se inicia a partir da ligação gratuita para o telefone 192, o sistema atende a urgências de natureza traumática, clínica, pediátrica, cirúrgica, gineco-obstetrica e mental da população. O atendimento de urgências e emergências pode ser realizado em residências, locais de trabalho e vias públicas (GHUSSN; SOUZA, 2016). A operação é coordenada com a rede de hospitais municipais, ou regionais, onde os hospitais e as estruturas de atendimento são escolhidos de acordo com suas especialidades (TAKEDA, 2000).

Incertezas são características básicas de SAE's e o SAMU não é exceção, como as relacionadas à demanda (REVELLE; EISELT, 2005), a disponibilidade dos servidores, o tempo de serviço, tempo de resposta dos servidores. Devido a essa característica dos SAE's, sem um adequado planejamento a qualidade dos sistemas pode ser comprometida (SOUZA, 2010). Em geral, a demanda e o tempo de serviço dos servidores podem ser modelados estatisticamente segundo alguma distribuição estatística (GALVÃO; MORABITO, 2008).

Esses desafios e características dos SAE's, os tornam objetos de estudo muito ricos, sendo observado um número crescente de estudos na área desde a década de 1950. Neste período, estudiosos do campo da pesquisa operacional desenvolveram seus trabalhos a partir de ferramentas como a programação matemática, estatística até teoria das filas (SIMPSON; HANCOCK, 2009). Dentre os problemas desenvolvidos neste período, pode-se citar os modelos de localização, responsáveis por otimizar a operação de SAE's a partir de decisões com respeito ao número e localização das unidades e suas políticas de despacho (CHIYOSHI *et al.*, 2001).

Os modelos de localização podem ser divididos em duas fases. Na primeira, emergiram modelos determinísticos; enquanto, na segunda, modelos probabilísticos mais realistas apareceram (RAJAGOPALAN *et al.*, 2008). Segundo Chiyoshi *et al.* (2000), os primeiros modelos desenvolvidos para a localização de SAE's foram determinísticos. Esses modelos buscam maximizar a cobertura de atendimento do sistema. Uma área é considerada coberta quando ela está dentro de uma área crítica, definida pelo tempo ou pela distância da base (SOUZA, 2010).

Os modelos probabilísticos levam em conta a natureza estocástica dos sistemas do mundo real. Eles consideram, por exemplo, a localização dos clientes, custos de construção, disponibilidade dos servidores como variáveis aleatórias. Encontram localizações ótimas para as instalações ou servidores com relação a uma medida de desempenho, considerando

explicitamente distribuições probabilísticas das variáveis modeladas (CHIYOSHI *et al.*, 2000). A incorporação das distribuições probabilísticas nas formulações dos problemas pode ser feita diretamente para a programação matemática, ou com um enfoque em teoria das filas (OWEN; DASKIN, 1998). O desenvolvimento dos modelos para servidores ocorreu juntamente com o desenvolvimento do modelo hipercubo de Larson (1974), ele é capaz de fornecer aos tomadores de decisão as medidas de desempenho para qualquer esquema de localização dos servidores (MARIANOV; REVELLE, 1996).

O modelo hipercubo é uma poderosa ferramenta que inclui a complexidade geográfica de uma cidade e políticas de despacho dos servidores. Como solução, o modelo calcula as probabilidades de estado do sistema analisado e, partir desse resultado, as medidas de desempenho associadas ao sistema. O nome do modelo deriva do espaço de estados que descreve a disponibilidade de todos servidores em livres ou ocupados (LARSON; ODONI, 2007).

Pode-se citar diversas aplicações do modelo hipercubo em SAE's pelo mundo, exaltando sua importância para os problemas de localização. Por exemplo, Brandeau e Larson (1986) em Boston, Davoudpour *et al.* (2014) em Teerã, Chelst e Barlach (1981) em New Haven, Geroliminis *et al.* (2009) em São Francisco, Geroliminis *et al.* (2011) em Atenas, entre vários outros. No Brasil, o modelo já foi utilizado em estudos em sistemas urbanos, como o SAMU em Campinas (TAKEDA, 2000; TAKEDA *et al.*, 2007), em Ribeirão Preto (SOUZA, 2010; SOUZA *et al.*, 2015), em Bauru (BEOJONE; SOUZA, 2017). Também foi utilizado para estudar SAE's em rodovias, como Iannoni *et al.* (2008, 2009), Iannoni (2005), Mendonça e Morabito (2001), Atkinson *et al.* (2006), entre outros.

No entanto, o modelo hipercubo não está livre de críticas. Mesmo após o desenvolvimento de alterações ao modelo básico, chamadas extensões do modelo, alguns problemas ainda persistem. Por exemplo, o modelo requer grande poder computacional quando o número de servidores aumenta. Larson (1975) e Jarvis (1985) desenvolveram modelos aproximados para trabalhar em sistemas muito grandes. Além disso, Chiyoshi *et al.* (2001, 2011), Larson e Odoni (2007), Morabito *et al.* (2008) Rodrigues *et al.* (2017) fazem menção a este problema.

Outros problemas do modelo hipercubo estão relacionados à maneira com que são representados os processos de serviço e os tempos de viagem. O modelo não considera a possibilidade de redespacho, onde um servidor é enviado para novo chamado antes de retornar à sua base (SOUZA *et al.*, 2015; RODRIGUES *et al.*, 2017). Ele também trabalha com localizações fixas, ao contrário de SAE's como os vistos em Alanis *et al.* (2013) e van

Barneveld *et al.* (2017), que reposicionam dinamicamente os servidores. A representação dos tempos de viagem também é problemática, especialmente quando esses tempos não são pequenos em relação ao tempo total de serviço, como em Geroliminis *et al.* (2011) e Boyaci e Geroliminis (2015). Outras críticas podem ser direcionadas ao despacho único (CHELST; BARLACH, 1981), às diferenças entre tempos de serviço (GEROLIMINIS *et al.*, 2009, 2011).

Muitos SAMU's trabalham classificando seus usuários em ordem de prioridades. Por exemplo, o SAMU de Campinas, estudado em Takeda (2000) e Takeda *et al.* (2004, 2007), classificava seus usuários em duas classes de chamados (avançados e básicos), mas sem considerá-las na fila de espera na construção do modelo hipercubo; o SAMU de Ribeirão Preto, estudo em Souza (2010) e Souza *et al.* (2015), separava seus usuários em três classes (emergência, urgência moderada e urgência leve) e considera a prioridade na fila de espera dentro do modelo. Para isso, foi necessário incorporar políticas de prioridade ao modelo hipercubo. Além da diferenciação dos usuários, os SAMU's diferenciam suas ambulâncias. Os chamados prioritários são atendidos preferencialmente pelos Veículos de Suporte Avançado (VSA's), enquanto outros tipos de chamados são atendidos pelos Veículos de Suporte Básico (VSB's).

O SAMU de Bauru, diferentemente dos modelos utilizados para os SAMU's de Campinas e de Ribeirão Preto, reserva seus VSA's para atender exclusivamente aos chamados prioritários. Assim, além de ser necessário modelar as diferentes prioridades de chamados na fila (SOUZA *et al.*, 2015), é preciso considerar certa reserva de capacidade (IANNONI *et al.*, 2015).

Ao mesmo tempo, ambulâncias numa mesma localização podem compartilhar os chamados. Além disso, o envio de uma ambulância para um chamado pode não seguir uma ordem predeterminada, pode ser aleatória. Assim, VSA's ou VSB's nessas condições costumam apresentar tempos de serviço iguais, o que os torna indistinguíveis (LUQUE, 2008).

Em outro sentido, a operação de serviços como SAE's requer que a capacidade de pessoal disponível combine com a demanda pelo serviço ao longo do dia. Esse tipo de problema deve ser resolvido determinando quando e quantos usuários chegam ao sistema (CHUNG; MIN, 2014). O desafio do gestor é programar os horários de serviço de forma que acompanhe a demanda em diferentes horários do dia, enquanto mantém-se os custos controlados e respeitando todas legislações aplicáveis (INGOLFSSON *et al.*, 2002). O requisito fundamental é que haja pessoal suficiente trabalhando para atingir níveis de serviço planejados (GREEN *et al.*, 2001).

A programação (ou *scheduling* em inglês) dos servidores é um desafio para qualquer serviço, como bancos, restaurantes, lojas, aeroportos e *call-centers* são apenas alguns dos exemplos (INGOLFSSON *et al.*, 2010). *Call-centers* talvez sejam a operação mais estudada para este tipo de problema, visto que podem fazer parte da operação de atendimento ao consumidor, de central de ajuda e de SAE's (GANS *et al.*, 2003). Assim como os SAE's, a operação desses sistemas está sujeita a incertezas quanto à demanda, e disponibilidade de pessoal (MANDELBAUM *et al.*, 2009; PATRICK *et al.*, 2008).

Uma forma simples de representar o problema de programação dos servidores é através de uma sequência de passos (INGOLFSSON *et al.*, 2002). No primeiro passo faz-se uma previsão da demanda período a período para o sistema. No segundo passo converte-se as previsões da demanda em requisitos de pessoal. No terceiro passo determina-se o conjunto de turnos possíveis. Por fim, no quarto passo, encontra-se um programa que obedeça a todos os requisitos e minimize uma função objetivo, como número de empregados, ou custos totais (BUFFA *et al.*, 1976).

As previsões de demanda são realizadas, usualmente, a partir de dados históricos (GILLARD; KNIGHT, 2014). A demanda possui uma variabilidade previsível quanto à média de chegadas ao longo do dia; e uma variabilidade estocástica sobre a média previsível causada pelo comportamento dos usuários e servidores (WHITT, 2007). Nesse contexto, espera-se que os processos de chegada sejam bem modelados por um processo de Poisson não-homogêneo (KIM; WHITT, 2014).

A obtenção dos requisitos de pessoal é normalmente obtida por uma sequência de modelos de filas simples, de forma independente, para cada período de tempo. Green *et al.* (2001) chama essa abordagem de "estacionário independente período a período" (SIPP – *Stationary Independent Period by Period*, em inglês). No entanto, o SIPP requer que a capacidade de atendimento do sistema seja sempre superior à demanda, não é permitida sobrecarga no sistema, mesmo que em períodos pequenos (STOLLETZ, 2008).

Devido à natureza não linear dos problemas de programação de servidores, a geração de programas é usualmente resolvida por meio de heurísticas. Por exemplo, o algoritmo genético (INGOLFSSON *et al.*, 2002) e a heurística utilizada em Chung e Min (2014).

Os modelos estacionários não conseguem capturar o comportamento do sistema nos momentos em que os servidores estão programados para terminarem seus turnos, ou pararem para uma refeição. Neste caso, existem duas possibilidades: o servidor interrompe o serviço (disciplina preemptiva), ou o servidor finaliza o atendimento antes de parar (disciplina não-preemptiva – chamada de exaustiva ao longo do texto). A disciplina preemptiva é a mais

utilizada na literatura, por sua simplicidade computacional, muito embora não seja a mais adequada quando os usuários do sistema foram pessoas. Por outro lado, a disciplina exaustiva, formulada analiticamente em Ingolfsson (2005), costuma ser mais realista quando os usuários forem humanos (INGOLFSSON *et al.*, 2007).

A representação do comportamento exato do sistema é feita por meio das equações diferenciais de Chapman-Kolmogorov (INGOLFSSON *et al.*, 2002). Essas equações não fornecem os requisitos de servidores para o segundo passo da programação de servidores, mas mostram se o objetivo de desempenho foi atingido, quando resolvidas numericamente. Elas conseguem capturar o comportamento das mudanças de servidores por meio de modelos de cadeias de Markov mistas discretas e contínuas (GILLARD; KNIGHT, 2014).

As equações de Chapman-Kolmogorov possuem soluções analíticas apenas em casos especiais (quando o sistema possui infinitos servidores e as funções das taxas de chegada e de serviço são bem-comportadas) e, portanto, devem ser resolvidas numericamente pelos métodos de Euler ou Runge-Kutta. Dessa maneira, embora precisas, a resolução numérica das equações toma um longo tempo computacional (SCHWARZ *et al.*, 2016). Algumas aproximações foram desenvolvidas como a uniformização (ou randomização) utilizada em Ingolfsson *et al.* (2007), que transforma uma cadeia de Markov de tempo contínuo em uma cadeia de Markov de tempo discreto (GRASSMANN, 1977); o atraso continuado estacionário, utilizada em Stolletz e Lagershausen (2013), que utiliza modelos estacionários e distribui o atraso de um período para o próximo (STOLLETZ, 2008); entre outras. No entanto, em algumas situações, as aproximações estacionárias não costumam apresentar bons resultados, um exemplo disso ocorre quando os tempos de serviço são longos (GREEN *et al.*, 2007). Além disso, as aproximações não costumam ser boas em sistemas em que o intervalo de tempo até o próximo evento (chegada ou término de um serviço) seja longo (INGOLFSSON *et al.*, 2002). O SAMU é um caso em que os tempos de serviço são longos e o número de eventos é pequeno, sendo mais indicado o uso das equações de Chapman-Kolmogorov.

Durante a operação diária de um SAMU podem haver variações significativas quanto à demanda em termos de número de chamados e distribuição geográfica (SOUZA, 2010). O estudo do SAMU de Ribeirão Preto considerou o dimensionamento em mais de um período do dia (SOUZA *et al.*, 2013). Contudo, este estudo não considerou questões de dimensionamento de pessoal e demanda considerando eventos diários como a trocas de turno, pausas para refeições e variações rápidas na demanda. No mesmo sentido, outro estudo foi realizado em Rajagopalan *et al.* (2008), considerando o modelo hipercubo aproximado em um SAE em três períodos diferentes.

No contexto levantado até aqui, uma pesquisa promissora para a análise de SAMU's parece seguir em duas direções quase paralelas: melhoria da configuração do sistema e avaliação do desempenho ao longo do tempo.

A primeira direção é o desenvolvimento de uma abordagem que utilize o modelo hipercubo no apoio à tomada de decisão quanto às ambulâncias, as bases e suas configurações. Nesse sentido, a verificação da relação entre o ganho e o ônus gerados a partir do crescimento tanto da cidade, quanto do SAMU se torna fundamental. Ainda pode-se levar em conta a relação entre a distribuição da carga de trabalho e o tempo médio de resposta, fundamentais para uma boa operação de um SAE.

A segunda direção também é o desenvolvimento de uma abordagem que utilize o modelo hipercubo, porém considerando as variações temporais pelas quais o sistema passa ao longo do dia. Assim, pode-se verificar o quão bem foi escolhido o momento para a troca de turnos em relação às cargas de trabalho. Assim como, a melhoria do rearranjo das pequenas pausas ao longo do dia, como as refeições, especialmente em circunstâncias adversas.

Nesse contexto é importante incorporar a reserva de capacidade de seus VSA's para o atendimento de chamados prioritários. Essa análise deve garantir um bom uso dos recursos computacionais, buscando eficiência, em termos da velocidade de processamento, e eficácia, em termos da qualidade dos resultados gerados. Além disso, é necessário incorporar a diferenciação dos servidores enquanto descreve o comportamento do sistema ao longo do dia de operação, sem esquecer de garantir a eficiência e eficácia do modelo. Foi desenvolvido um estudo de caso no SAMU de Bauru para verificar a aplicabilidade dessa abordagem em um sistema real.

1.1 Objetivos do trabalho

Dessa maneira, este trabalho possui como objetivo geral propor e avaliar abordagens baseadas no modelo hipercubo de filas para análise, planejamento e, futura, otimização da configuração e operação de SAE's urbanos ao longo do dia, de forma a mostrar a importância de se considerar o comportamento não-estacionário dos sistemas. Pretende-se realizar a avaliação do desempenho e de cenários alternativos do SAMU de Bauru, utilizando as abordagens desenvolvidas para mostrar que elas são efetivas para a análise desse tipo de sistema, assim como destacar as diferenças obtidas entre o uso de modelos estacionários e não-estacionários. Os objetivos específicos, aplicados em um estudo de caso no SAMU de Bauru, são descritos a seguir:

i) Estender o modelo hipercubo de filas para que seja capaz de capturar a reserva total de capacidade das ambulâncias avançadas (VSA's). A reserva de capacidade deve manter-se tanto em estados de fila, quanto sem fila, como forma de garantir um atendimento adequado especialmente para os chamados com risco de vida;
ii) Estender o modelo hipercubo de filas no sentido de torná-lo eficiente computacionalmente, sem haver perda adicional (em relação do modelo hipercubo) de precisão durante a modelagem e resolução; e
iii) Propor uma abordagem baseada no modelo hipercubo para organização do trabalho das ambulâncias em qualquer momento do dia, o modelo deve ter comportamento não-estacionário. Devido às características do SAMU de Bauru, como seus tempos de serviço longos e se pequeno número de eventos, a modelagem precisa ser feita a partir das equações de Chapman-Kolmogorov e resolvidos numericamente pelo método de Runge-Kutta.

As propostas de abordagens deste trabalho contribuem para a gestão de SAE's (especialmente de SAMU's), tanto para o gerenciamento estratégico-tático da operação, quanto para nível operacional. O gerenciamento estratégico-tático é beneficiado pelo desenvolvimento das extensões mais precisas para a localização e organização das ambulâncias e suas bases. Do ponto de vista operacional, o benefício ocorre pelo desenvolvimento de um modelo capaz de organizar e captar o comportamento das equipes. Os benefícios ocorrem e são mensurados por meio de cenários alternativos e do cálculo de medidas de desempenho como carga de trabalho, nível de serviço, tempo médio de espera, tempo médio de resposta, etc.

1.2 Método de pesquisa

A pesquisa realizada pode ser classificada como modelagem e simulação. Por seguir um estudo de caso, tentando entendê-lo melhor utilizando-se de teoria das filas e propor melhorias ao sistema real estudado, através do estudo de cenários alternativos, sua tipologia é a pesquisa empírica descritiva e, também, normativa.

1.3 Estrutura do trabalho

Os próximos Capítulos deste trabalho foram organizados como segue.

O Capítulo 2 traz a revisão do modelo hipercubo e suas extensões já existentes para a análise de sistemas espacialmente distribuídos. O capítulo foca nas extensões mais relacionadas

e utilizadas ao longo do trabalho, mas também apresenta outras já vistas na literatura. Assim como visita trabalhos em que tais modelos foram utilizados dentro de um modelo de otimização.

O Capítulo 3 traz uma revisão dos modelos de filas encontrados que não abordam sistemas espacialmente distribuídos, e que lidam com parâmetros variáveis no tempo. Foram separados os estudos que utilizam métodos aproximados a partir dos modelos estacionários dos estudos com modelos exatos. Os modelos exatos também foram explicados com mais detalhes quanto às mudanças de turno.

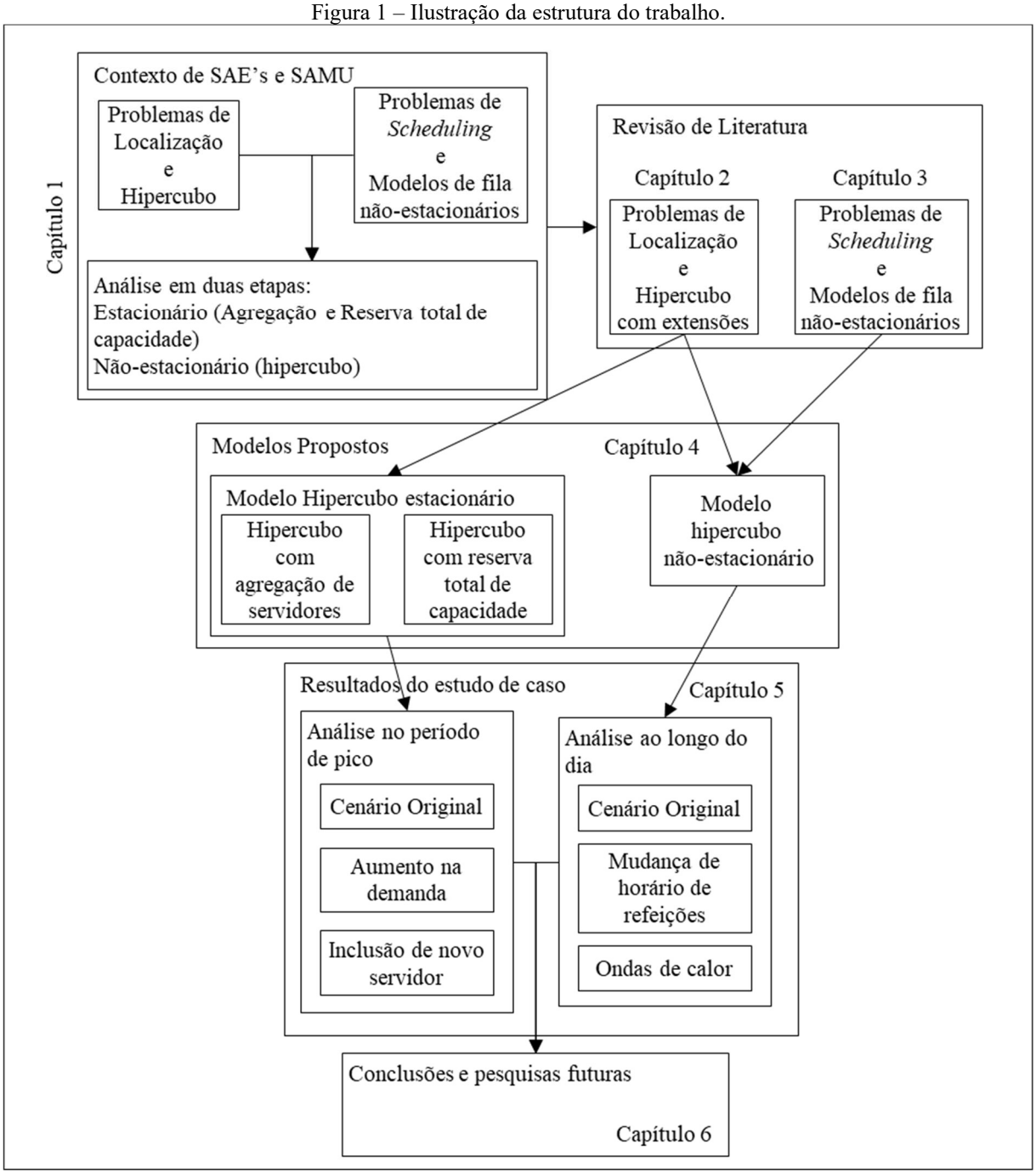
O Capítulo 4 apresenta as extensões desenvolvidas a partir das características do SAMU-Bauru. Além das extensões, também se apresenta a maneira como as políticas de despacho foram traduzidas ao longo do estudo de caso. As extensões são apresentadas por meio de exemplos ilustrativos. A primeira extensão trata da reserva de capacidade das ambulâncias avançadas do SAMU-Bauru. A segunda extensão trata da agregação de servidores para diminuição do espaço de estados. Por fim, apresenta-se a extensão para o estudo do modelo hipercubo não-estacionário, assim como suas possíveis disciplinas de fim de turno.

O Capítulo 5 faz uma apresentação do SAMU-Bauru e dos resultados computacionais obtidos durante o estudo de caso. A apresentação do SAMU-Bauru mostra a configuração do sistema, como o processo de atendimento, a política de despacho, a localização das bases e das ambulâncias, etc. Nesse capítulo também são apresentados os resultados do estudo em duas etapas, primeiro o estudo estacionário, seguido pelo não-estacionário. O estudo estacionário apresenta uma comparação do modelo utilizado com a amostra e a avaliação do desempenho quando o sistema passa por aumentos na demanda e a inclusão de uma nova ambulância. O estudo não-estacionário compara a análise feita utilizando modelos estacionários e também avalia mudanças nos horários das refeições e o aumento na demanda em horários específicos do dia.

Finalmente, no Capítulo 6 são apresentadas as conclusões finais do trabalho como um resumo dos principais resultados encontrados, as contribuições realizadas, as limitações e as perspectivas de pesquisas futuras.

A Figura 1 ilustra a estrutura criada para o trabalho e as interações entre seus capítulos.

Figura 1 – Ilustração da estrutura do trabalho.



2 MODELO HIPERCUBO E SUAS EXTENSÕES

O modelo hipercubo possibilita a utilização de filas espacialmente distribuídas em modelos de localização probabilística. Por ser um modelo descritivo, ele não encontra por si só uma solução para um problema de localização. No entanto, por medir o desempenho e a interação entre os servidores, ele é utilizado por alguns autores em métodos heurísticos para resolver problemas de localização probabilísticos (CHIYOSHI *et al.*, 2000).

Daskin (1983) foi um trabalho pioneiro ao estender os modelos de localização, até então determinísticos, considerando a possibilidade de que servidores possam estar ocupados e indisponíveis. Seu modelo, chamado de problema de localização de máxima cobertura esperada (MEXCLP – *Maximum Expected Covering Location Problem*, em inglês), busca maximizar a cobertura esperada considerando que uma probabilidade de os servidores estarem ocupados.

De forma alternativa, ReVelle e Hogan (1989) propôs o problema de localização de máxima disponibilidade (MALP – *Maximum Availability Location Problem*, em inglês). O objetivo desse modelo é maximizar a população coberta por um servidor disponível dentro de um tempo de resposta estabelecido dentro de um nível de confiabilidade.

Os modelos MEXCLP e MALP serviram como base para boa parte de outros estudos mais recentes que flexibilizam, basicamente, suas três hipóteses simplificadoras. A primeira hipótese é a independência dos servidores. A segunda hipótese é a taxa de ocupação idêntica para todos servidores. Por fim, a terceira hipótese é que a taxa de ocupação dos servidores não varia conforme suas respectivas localizações (IANNONI, 2005). Por esse motivo, o modelo hipercubo se mostra como forma adequada para tratar esses problemas de localização, já que não possui essas hipóteses simplificadoras.

Batta *et al.* (1989) revisitou o modelo MEXCLP propondo um método iterativo integrado ao modelo hipercubo como forma de relaxar as hipóteses simplificadoras do MEXCLP. O método proposto é um procedimento heurístico em que uma solução inicial tem a localização de um único servidor alterada. A alteração é feita até se encontrar uma solução melhor que a anterior, ou não ser possível encontrar nenhuma solução melhor substituindo apenas a localização de um único servidor. A solução deve obter a maior cobertura esperada calculada a partir do resultado do modelo hipercubo. Neste trabalho também foi apresentado o modelo AMEXCLP, baseado no modelo hipercubo aproximado de Larson (1975).

Por um lado, Saydam *et al.* (1994) avaliou a acuracidade do modelo AMEXCLP, reiterando a importância de se utilizar o modelo hipercubo em modelos de otimização. Enquanto por outro, Chiyoshi *et al.* (2003) investigou as diferenças obtidas entre os modelos

MEXCLP, AMEXCLP e procedimento iterativo baseado no modelo hipercubo de Batta *et al.* (1989), chamado de HLM (*Hypercube Location Model*, em inglês). O HLM se mostrou o mais promissor por permitir trabalhar com chamados em fila.

Galvão *et al.* (2003) propõe um método heurístico para estender o modelo MALP. Semelhantemente a Batta *et al.* (1989), é proposto um processo iterativo de substituição da localização de um servidor para encontrar soluções melhores do que a inicial, utilizando o modelo hipercubo para a avaliação. Sendo que este modelo também estendeu o MALP para que o servidor localizado em um nó seja conhecido. Por isso, esse modelo foi chamado de EMALP (*Extended* MALP, em inglês).

Iannoni (2005) e Iannoni *et al.* (2008) estudaram a combinação do modelo hipercubo com o algoritmo genético como forma de otimizar a configuração de SAE's em rodovias. O problema buscou minimizar os tempos médios de resposta e/ou balancear as cargas de trabalho determinando, por exemplo, o tamanho da área de cobertura das ambulâncias.

Geroliminis *et al.* (2009) propôs um modelo para minimizar o tempo médio de resposta considerando uma cobertura fixa, com uma escolha de envio de servidores ótima. Para resolver o problema foi utilizado um problema de programação não-linear. Por outro lado, Geroliminis *et al.* (2011) aplicou o algoritmo genético para a solução do problema.

Boyaci e Geroliminis (2015) utilizam o modelo hipercubo juntamente de um algoritmo de partição para agrupar servidores de forma a minimizar a interação entre os grupos. É feito o uso de busca local e arrefecimento simulado para encontrar soluções próximas a um ótimo. Seus resultados foram comparados ao MEXCLP como forma de mostrar a aplicabilidade dos modelos.

Rajagopalan *et al.* (2008) apresentou um modelo que visa determinar o número mínimo de ambulâncias e suas localizações, enquanto atinge um nível de cobertura. Esse modelo foi aplicado para situações em que há grande variação no comportamento da demanda.

Podemos citar ainda alguns estudos que buscaram estudar os modelos já apresentados utilizando diferentes heurísticas para encontrar a solução. Por exemplo, Saydam e Aytug (2003) utiliza o algoritmo genético para o MEXCLP, enquanto Galvão *et al.* (2005) propôs o uso de arrefecimento simulado para resolver o AMEXCLP e o EMALP.

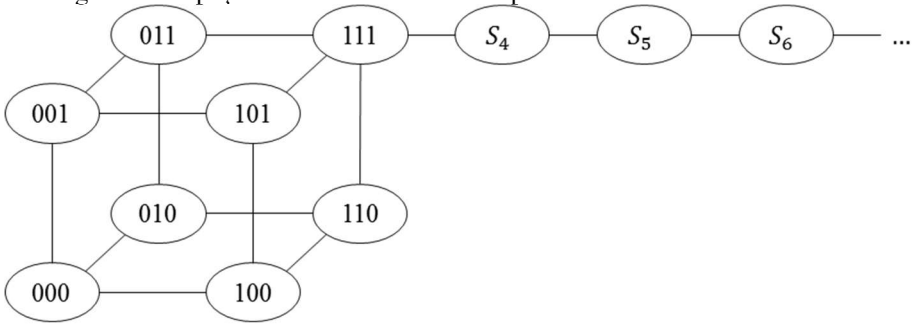
Outros estudos alteraram de alguma forma os modelos já conhecidos. Erkut *et al.* (2007) buscou maximizar a sobrevivência em modelos como o MEXCLP, no lugar da cobertura. Ingolfsson *et al.* (2008) buscou minimizar o número de ambulâncias necessárias para se atingir um nível de serviço determinado. Davoudpour *et al.* (2014) modifica o MEXCLP para maximizar a cobertura considerando disponibilidade e preferência de despacho.

2.1 O modelo hipercubo

Desenvolvido por Larson (1974), o modelo hipercubo de filas é um modelo analítico, baseado em sistemas de filas espacialmente distribuídas. A ideia é fazer uma expansão dos estados de um sistema *M/M/s*, buscando representar os *s* servidores individualmente em um sistema *server-to-customer* (servidor vai até o usuário). O modelo permite trabalhar com políticas de despacho mais complicadas. Em sua forma original, resolver o modelo significa encontrar a solução do sistema, para isso é preciso elaborar e resolver um conjunto de equações de equilíbrio (*steady state*), os resultados são as probabilidades de ocorrência dos estados do sistema em equilíbrio. A partir da solução do modelo podem ser calculadas diversas medidas de desempenho para o sistema, tais como: carga de trabalho dos servidores, frequências de despacho de servidores, tempos médios de viagem, tempos médios de fila, entre outras (SOUZA, 2010).

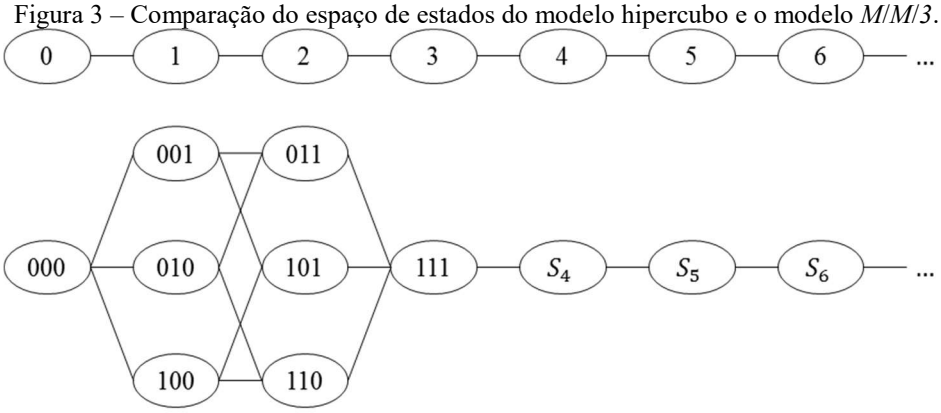
A disponibilidade dos servidores é representada por meio do espaço de estados do sistema. Por exemplo, um estado particular do sistema é dado pela lista dos servidores que estão livres ou ocupados. Considere um sistema com $N = 3$ servidores e sejam $\{000\}, \{001\}, \{010\}, \ldots, \{111\}$ os $2^3 = 8$ possíveis estados do sistema. O número "0" indica um servidor livre, enquanto o número "1" indica um servidor ocupado. Por exemplo, o estado $\{011\}$ representa a situação em que o servidor 1 está livre e os servidores 2 e 3 estão ocupados (CHIYOSHI *et al.*, 2001). O espaço de estados desse sistema com três servidores pode ser representado por um cubo; caso haja mais do que três servidores, tem-se um hipercubo. A Figura 2 ilustra o espaço de estados desse sistema com três servidores. Os estados $S_4, S_5, S_6, \ldots$ representam os estados com 1, 2, 3, ..., usuários em fila de espera, respectivamente, atendidos segundo uma disciplina FCFS (*First Come First Served*, em inglês). O modelo hipercubo trata tanto sistemas que aceitam a formação de fila quanto sistemas que não aceitam.

Figura 2 – Espaço de estados do modelo hipercubo com três servidores e fila.



Geroliminis *et al.* (2009) utiliza uma representação do espaço de estados que permite visualizar mais facilmente a relação do modelo hipercubo com o modelo *M*/*M*/*3*. A Figura 3 mostra esta representação para um sistema com três servidores. Note que para cada estado do modelo *M*/*M*/*s* existe um conjunto de estados no modelo hipercubo com a mesma quantidade de usuários no sistema.

Figura 3 – Comparação do espaço de estados do modelo hipercubo e o modelo *M*/*M*/*3*.



2.2 Modelo hipercubo clássico

Segundo Larson e Odoni (2007), a aplicação do modelo hipercubo requer a verificação das nove hipóteses apresentadas a seguir:

i) Existência de átomos geográficos: a região onde são prestados os serviços do sistema deve ser dividida em N_A átomos geográficos, sendo que cada átomo corresponde a uma fonte independente de chamados;

ii) Processo de chegada: deve ser um processo de Poisson homogêneo. Os usuários de cada átomo solicitam chamados por meio do processo de Poisson, sendo os chamados independentes entre si. Além disso, as taxas de chegada, λ_j, de chamados de cada átomo deve ser conhecida;

iii) Tempos de viagem dos servidores aos átomos: Os tempos de viagem, τ_{ij}, de cada servidor i para o átomo j devem ser conhecidos ou estimados;

iv) Servidores do sistema: existem *N* servidores espacialmente distribuídos ao longo do sistema, sendo que cada um pode se deslocar e atender a qualquer um dos átomos;

v) Localização dos servidores: A localização do servidor no sistema deve ser conhecida ao menos probabilisticamente. O servidor pode se mover pelos átomos, localização probabilística, ou ficar fixo em um deles;

vi) Despachos dos servidores: para atender a qualquer chamado é enviado apenas um servidor para o local. Se não houverem servidores disponíveis os chamados entrarão em fila ou serão considerados perdas do sistema;

vii) Política de despacho dos servidores: há uma lista de preferência de despacho para cada átomo, ou seja, deve ser obedecida uma ordem de envio dos servidores para os chamados;

viii) Tempo de serviço: o tempo de serviço de um servidor engloba o tempo de *setup*, o tempo de viagem e o tempo em cena até o retorno à base (ou área) de origem; e

ix) Dependência do tempo de serviço em relação ao tempo de viagem: a variação do tempo de viagem deve ser considerada uma variável de segunda ordem no tempo total de serviço, quando comparado ao tempo em cena e preparação da equipe. Isso não quer dizer que o tempo de viagem seja ignorado no cálculo do tempo médio de serviço, já que este é incorporado através da calibração do tempo médio de serviço (μ^{-1}), onde μ^{-1} é igual à soma do tempo médio de viagem do servidor e do tempo médio de atendimento.

A seguir é apresentado um exemplo de aplicação do modelo hipercubo clássico encontrado em Chiyoshi *et al.* (2000). Considere um sistema, que não admite fila de espera, cuja região é dividida em três átomos que são atendidos conforme a matriz de preferência de despacho fixa presente na Tabela 1.

Tabela 1 – Matriz de preferência de despacho para exemplo do modelo hipercubo clássico.

Átomo	Preferência		
	1º	2º	3º
1	1	2	3
2	2	3	1
3	3	1	2

Fonte: Chiyoshi *et al.* (2000, p148) adaptado.

Para encontrar a solução do sistema é preciso elaborar e resolver um conjunto de equações em equilíbrio (*steady state*) elaboradas a partir de cada estado do sistema. Esse sistema pode ser resolvido como um sistema linear homogêneo determinado caso uma de suas equações de equilíbrio seja substituída pela equação $\sum_{B \in D} P_B = 1$ que mostra que a soma das probabilidades do sistema é igual à 1. Os resultados são as probabilidades de estado do sistema em equilíbrio. A Equação (1) representa tal conjunto para o sistema considerado. Em que μ_i é a taxa de serviço do servidor *i* (μ é a taxa de serviço total) e λ_j é a taxa de chegada do átomo *j* (λ é a taxa de chegada total). Note que devido à matriz de preferência de despacho a transição do estado {100} para o estado {110} possui uma taxa de $\lambda_1 + \lambda_2$ já que o servidor preferencial

do átomo 1, o servidor 1, está ocupado e a próxima opção para envio é o servidor 2. Soma-se a isso a taxa de chegada do átomo 2, onde o servidor 2 é o preferencial.

$$\begin{aligned}
&\lambda P_0 = \mu_1 P_{\{100\}} + \mu_2 P_{\{010\}} + \mu_3 P_{\{001\}} \\
&(\lambda + \mu_1) P_{\{100\}} = \mu_2 P_{\{110\}} + \mu_3 P_{\{101\}} + \lambda_1 P_{\{000\}} \\
&(\lambda + \mu_2) P_{\{010\}} = \mu_3 P_{\{011\}} + \mu_1 P_{\{110\}} + \lambda_2 P_{\{000\}} \\
&(\lambda + \mu_3) P_{\{001\}} = \mu_2 P_{\{011\}} + \mu_1 P_{\{101\}} + \lambda_3 P_{\{000\}} \\
&(\lambda + \mu_1 + \mu_2) P_{\{110\}} = \mu_3 P_{\{111\}} + (\lambda_1 + \lambda_2) P_{\{100\}} + \lambda_1 P_{\{010\}} \\
&(\lambda + \mu_1 + \mu_3) P_{\{101\}} = \mu_2 P_{\{111\}} + \lambda_3 P_{\{100\}} + (\lambda_1 + \lambda_3) P_{\{001\}} \\
&(\lambda + \mu_2 + \mu_3) P_{\{011\}} = \mu_1 P_{\{111\}} + (\lambda_2 + \lambda_3) P_{\{010\}} + \lambda_2 P_{\{001\}} \\
&(\lambda + \mu) P_{\{111\}} = \lambda \left(P_{\{110\}} + P_{\{101\}} + P_{\{011\}} \right)
\end{aligned} \tag{1}$$

A partir das probabilidades de estado pode-se calcular diversas medidas de desempenho para o sistema. Elas podem ser separadas em internas e externas. As medidas internas são aquelas observadas do ponto de vista do sistema, do gestor, sendo elas: carga de trabalho dos servidores, frequências de despacho de servidores, tempos médios de viagem, etc. As medidas externas são as observadas a partir do ponto de vista do usuário, como: tempos médios de fila, tempos médios de resposta, entre outras.

A carga de trabalho (*workload*) de um servidor do sistema é a fração de tempo em que o servidor está ocupado. Para calcular essa importante medida de desempenho é necessário somar as probabilidades de o servidor estar ocupado, as probabilidades de estado em que estão em serviço $\{1\}$ mais as probabilidades dos estados em fila, em que todos servidores já estão ocupados (Equação (2)).

$$\rho_i = \sum_{B \in E_i} P_B \tag{2}$$

Em que E_i é o conjunto de estados do sistema onde o servidor i está ocupado.

Outra medida de desempenho importante é a frequência de despacho. Ela é necessária para se observar a utilização de *backups* e calcular outras medidas de desempenho. O seu cálculo envolve os despachos realizados sem espera (nq) e com espera (q), conforme mostra a Equação (3).

$$f_{ij} = f_{ij}^{(nq)} + f_{ij}^{(q)} = \frac{\lambda_j}{\lambda} \sum_{B \in E_{ij}} P_B + \frac{\lambda_j}{\lambda} P_Q' \frac{\mu_i}{\mu} \tag{3}$$

Em que:

- λ_j / λ é a fração das chegadas correspondentes ao átomo *j* no sistema;

- $\sum_{B \in E_{nj}} P_B$ é a soma das probabilidades de envio do servidor *i* ao átomo *j*, quando o servidor *i* estiver livre e for o próximo a ser enviado ao átomo *j* pela matriz de preferência de despacho;
- P'_Q é a probabilidade de saturação do sistema, a soma da probabilidade dos estados de fila mais a probabilidade de todos servidores estarem ocupados; e
- μ_i/μ é a probabilidade de o servidor *i* ser o primeiro liberado, quando todos servidores estiverem ocupados.

Outro conjunto de medidas de desempenho central para o sistema são os tempos de viagem. Eles são calculados a partir da matriz de tempos de viagem (τ) e da matriz de localização (l). O tempo médio de um servidor *i* viajar ao átomo *j*, quando disponível, conforme a Equação (4).

$$t_{ij} = \sum_{k=1}^{N_A} l_{ik}\tau_{kj} \tag{4}$$

A partir da matriz de tempos de viagem, pode-se calcular os tempos médios de viagem para chamados sujeitos à espera em fila. Esse cálculo é feito a partir das proporções de um chamado chegar no átomo *j* e de um servidor ser despachado do átomo *i*, conforme visto na Equação (5).

$$\bar{T}_Q \equiv \sum_{i=1}^{N_A}\sum_{j=1}^{N_A} \frac{\lambda_i \lambda_j}{\lambda^2}\tau_{ij} \tag{5}$$

O tempo médio de viagem do sistema, independentemente do servidor e do átomo pode ser calculado pela Equação (6).

$$\bar{T} = \sum_{i=1}^{N}\sum_{j=1}^{N_A} f_{ij}^{(nq)} t_{ij} + P'_Q \bar{T}_Q \tag{6}$$

O tempo médio de viagem ao átomo *j*, independentemente do servidor designado é dado pela Equação (7).

$$\bar{T}_j = \frac{\sum_{i=1}^{N} f_{ij}^{(nq)} t_{ij}}{\sum_{i=1}^{N} f_{ij}^{(nq)}}\left(1 - P'_Q\right) + \sum_{k=1}^{N_A}\left(\frac{\lambda_k}{\lambda}\right)\tau_{kj} P'_Q \tag{7}$$

Os tempos médios de viagem do servidor *i*, independentemente do átomo de destino, é dado pela Equação (8).

$$\overline{TU}_i = \frac{\sum_{j=1}^{N_A} f_{ij}^{(nq)} t_{ij} + \frac{\mu_i}{\mu}\bar{T}_Q P'_Q}{\sum_{i=1}^{N} f_{ij}^{(nq)} + \frac{\mu_i}{\mu} P'_Q} \tag{8}$$

O tempo de espera do sistema pode ser calculado por meio da Fórmula de Little, conforme indicado na Equação (9) (LITTLE, 2011).

$$W_Q = \frac{L_Q}{\lambda(1 - P_{perda})} \tag{9}$$

Outras medidas de desempenho também podem ser calculadas, por exemplo a fração de chamados atendidos por servidores de *backup*, entre outras (LARSON; ODONI, 2007).

As aplicações do modelo podem ser encontradas na literatura. Larson (1974) apresenta o modelo e o aplica para o sistema de veículos de patrulha da polícia de Boston, trazendo opções para melhorar o serviço desse sistema. Jarvis (1985) traz um método aproximado para a resolução do modelo hipercubo aplicado para sistemas com servidores heterogêneos sem fila de espera, como os bombeiros. Mendonça e Morabito (2001) utilizam o modelo hipercubo para equilibrar a carga de trabalho dos servidores em um SAE rodoviário. Geroliminis *et al.* (2011) utiliza um modelo hipercubo híbrido com o uso de algoritmo genético para melhorar o despacho de servidores em sistemas urbanos de grande escala e dimensões geográficas.

Diversas dessas aplicações do modelo hipercubo em sistemas reais necessitam que o modelo sofra alterações. Embora possua hipóteses para sua aplicação, extensões do modelo clássico trabalham com alterações nessas hipóteses, algumas são discutidas a seguir.

2.3 Extensão 1: aleatoriedade no despacho

Uma das extensões exploradas na literatura é a inclusão de aleatoriedade no despacho dos servidores. Isto ocorre quando há uma matriz de preferência de despacho, contudo mais de um servidor possui a mesma preferência de envio, sendo um deles escolhido aleatoriamente para atender ao chamado. A experiência mostra que esta é uma característica razoavelmente comum, especialmente quando há mais de um servidor com a mesma localização (BURWELL *et al.*, 1993). Na literatura encontra-se duas formas de se representar sistemas com essa característica.

A primeira forma, utilizada em Takeda *et al.* (2007) e Burwell *et al.* (1993), resolve múltiplos modelos. A aleatoriedade é dada pela repetida solução do sistema utilizando preferências fixas e cálculo da média das probabilidades de cada estado ao final. A matriz de preferência de despacho é gerada aleatoriamente para cada solução do sistema e pode seguir regras pré-determinadas.

Esta forma possui uma série de problemas: não há representação formal da aleatoriedade, estando essa sujeita à algoritmos de geração de números aleatórios na criação

das matrizes de preferência de despacho, e também sendo necessário um grande número de repetições da solução do modelo para se obter uma solução satisfatória; o modelo hipercubo já possui um custo computacional muito elevado para sistemas com muitos servidores e a repetida solução do sistema aumenta ainda mais tal custo. As medidas de desempenho são calculadas conforme o modelo original em cada repetição e encontra-se a média para o resultado final de cada medida de desempenho.

Chiyoshi *et al.* (2011) faz uma representação formal da aleatoriedade, tornando-a possível de representar nas equações de equilíbrio, eliminando os problemas encontrados no método de Takeda *et al.* (2007) e Burwell *et al.* (1993). Contudo essa forma ainda possui dois problemas: o modelo continua com muitas equações de equilíbrio a serem resolvidas; como a representação da matriz de preferência de despacho é implícita nas equações do modelo, fica mais difícil representar políticas de despacho complexas. A chance de um servidor ser escolhido para atender a um chamado depende apenas dos servidores disponíveis. A Equação (10) mostra um exemplo de equação de equilíbrio para o estado $\{100\}$, um sistema com três servidores. Observa-se que a entrada no estado se dá pela chegada de um chamado em $\{000\}$, que é dividida em três $(\lambda/3)$, uma vez que o servidor 1 é escolhido aletoriamente entre os servidores disponíveis uma a cada três vezes para atender aos chamados. As outras entradas ocorrem a partir do término de um serviço dos outros servidores (2 e 3).

$$(\lambda + \mu_1)P_{\{100\}} = \frac{\lambda}{3}P_{\{000\}} + \mu_2 P_{\{110\}} + \mu_3 P_{\{101\}} \qquad (10)$$

Larson (1975) e Batta *et al.* (1989) mostram uma forma aproximada que utiliza fatores de correção para escolher os servidores enviados, a ideia é que seja enviado o servidor mais próximo de um chamado e a localização dos servidores é conhecida probabilisticamente. A probabilidade de um servidor ser enviado é proporcional ao produto da probabilidade deste servidor estar disponível pela probabilidade do servidor preferencial estar ocupado. Este processo não é exatamente aleatório, uma vez que os fatores de correção são na realidade determinísticos.

Nas aplicações em sistemas com aleatoriedade, Takeda *et al.* (2007) mostra o uso do modelo hipercubo para a elaboração de cenários para o SAMU-Campinas com o objetivo de melhorar os tempos de resposta aos usuários a partir da descentralização dos servidores disponíveis e pela análise de cenários alternativos, acrescentando novos.

2.4 Extensão 2: *backup* parcial

Segundo Iannoni *et al.* (2009), existem SAE's caracterizados por possuírem políticas de despacho particulares, em que pelo menos um átomo não é atendido por todos os servidores do sistema, chamadas de *backup* parcial. Dessa maneira, alguns átomos só podem ser atendidos por determinados servidores. Isso ocorre, em particular, em sistemas em que os tempos de viagem são longos, por exemplo, em rodovias. Caso ocorra um chamado e o servidor disponível mais próximo precise atravessar várias dezenas ou centenas de quilómetros é provável que o tempo de resposta seja insatisfatório e, em se tratando de SAE's, pode significar o aumento do risco de vida de um usuário. A chamada é, então, transferida para outro sistema, por exemplo o corpo de bombeiros, ou o SAMU mais próximo.

Há uma abordagem para esse tipo de sistema em Mendonça e Morabito (2001) quando foi realizada a avaliação do sistema Anjos do Asfalto. De maneira geral, SAE's em rodovias brasileiras não admitem fila de espera, como mostram os estudos de Mendonça e Morabito (2001) e Iannoni *et al.* (2008; 2009).

Por causa do *backup* parcial, o sistema terá: estados de não saturação, em que um novo chamado pode ser atendido pelo servidor disponível mais próximo; estados de saturação, em que qualquer novo chamado resultará em perda para o sistema; e estados de semi-saturação, em que a chegada de chamados dos átomos sujeitos ao *backup* parcial resulta em perda para o sistema, enquanto a chegada de chamados de outros átomos pode ser atendida de imediato (IANNONI; MORABITO, 2008). Devido aos estados de semi-saturação é preciso considerar a possibilidade de perda de usuários mesmo com servidores livres, as medidas de desempenho também têm alterações quanto ao modelo clássico.

Por haver formação de fila no sistema, a frequência de despacho do servidor i para o átomo j é calculado sem levar em conta os chamados em espera. Assim, é preciso considerar a probabilidade de perda do sistema no cálculo, como mostra a Equação (11). No caso de um servidor *i* que não pode atender ao átomo j, tem $f_{ij} = 0$, porque ele não é assinalado em nenhum caso como *backup* para o átomo j na matriz de preferência de despacho.

$$f_{ij} = \frac{\frac{\lambda_j}{\lambda} \sum_{B \in E_{ij}} P_B}{1 - P_{perda}} \tag{11}$$

Diferentemente do modelo *M*/*M*/*s*/*K* e do modelo hipercubo clássico, a probabilidade de perda do sistema não é mais igual à probabilidade de o sistema estar com todos os servidores ocupados, é preciso levar em conta a perda em estados de semi-saturação e no estado de saturação. No estado de saturação todos novos chamados são perdidos, nos estados de semi-

saturação uma fração do total da taxa de chegada de chamados é perdida, como mostra a Equação (12) (IANNONI; MORABITO, 2006).

$$P_{perda} = \sum_{E} \left(\frac{\sum_{S \epsilon T_E} \lambda_S}{\lambda} \cdot P_E \right) \quad (12)$$

Em que:

- T_E é o conjunto de átomos cujo chamado ocasionará em perda, caso esteja no estado E;
- $\sum_{S \epsilon T_E} \lambda_S / \lambda$ é a fração de chamados perdidos no estado E;

Mendonça e Morabito (2001) utiliza o modelo com *backup* parcial no sistema Anjos do Asfalto e faz uma avaliação das consequências de aumento na demanda. Também é feito um balanceamento das cargas de trabalho ajustando o tamanho dos átomos do sistema.

Iannoni *et al.* (2009) aplica o modelo com *backup* parcial para otimizar o atendimento de um SAE em rodovia utilizando algoritmo genético. O método consegue combinar duas decisões a respeito de SAE's em rodovias, a localização das ambulâncias e o tamanho dos átomos. Ele é aplicado à dois SAE's em estudos de caso e seus resultados são posteriormente validados por meio de simulação de eventos discretos.

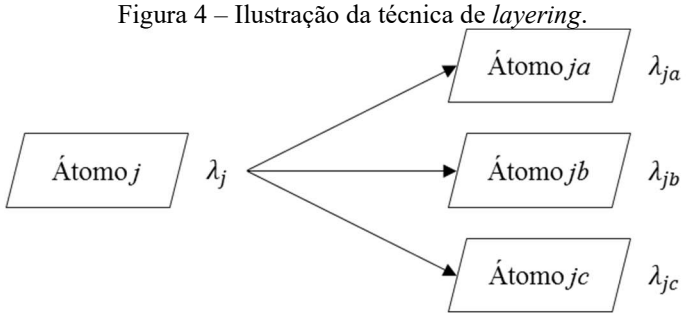
2.5 Extensão 3: Prioridade em fila

Um sistema emergencial urbano, como o SAMU, é composto por vários tipos de servidores (ambulâncias) (TAKEDA *et al.*, 2007). As ambulâncias podem ser classificadas em avançadas (VSA's) ou básicas (VSB's). Não é interessante enviar um VSA para atender chamados mais simples, sem risco de vida, visto que estas ambulâncias são minorias nos sistemas e sua operação é mais cara, já que são compostas por médicos e enfermeiros e possuem equipamentos especializados para atender chamados com risco de vida (TAKEDA *et al.*, 2007; SOUZA *et al.*, 2015). Por outro lado, as VSB's são mais frequentes e de menor custo, já que é operada com técnicos em enfermagem e os motoristas e têm equipamentos mais básicos para atender chamados que não incorrerem risco de vida eminente.

Takeda *et al.* (2007) apresentou uma forma de separar os chamados conforme sua gravidade por meio de uma técnica denominada *layering*, essa técnica consiste em separar os átomos geográficos em camadas (subátomos), cada camada é tratada como um átomo e possuem fontes de usuários independentes e preferência de despacho. Nas listas de preferência de despacho, os subátomos de maior nível de emergência (prioridade) possuem os VSA's como seus servidores prioritários acompanhados dos VSB's como *backups*. Por outro lado, os

subátomos de menor prioridade possuem os VSB's como prioritários e os VSA's como última alternativa, partindo do princípio de que o sistema não se considera o *backup* parcial. A Figura 4 ilustra o processo de *layering* de um átomo *j* em três subátomos *a*, *b* e *c*, em que *a* possui prioridade maior do que *b* que por sua vez é maior do que *c*. No trabalho de Takeda *et al.* (2007) não houve diferenciação dos chamados presentes na fila, que seguiam uma disciplina FCFS.

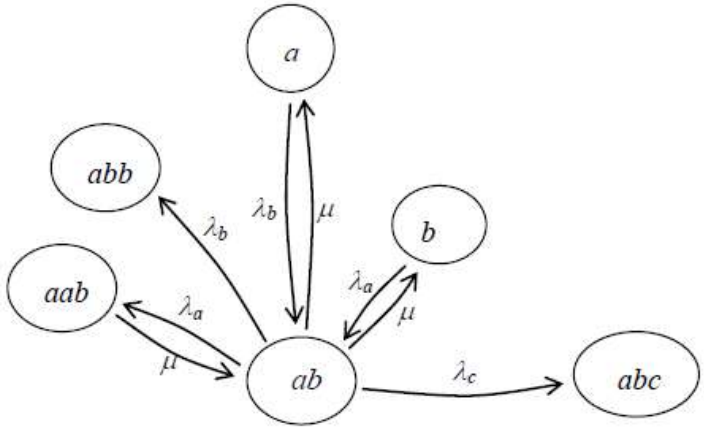
Figura 4 – Ilustração da técnica de *layering*.



Souza *et al.* (2015) apresenta uma forma de representação dos estados da fila com suas prioridades, a denominação das prioridades segue a nomenclatura dada aos subátomos. A prioridade máxima é denominada *a*, seguida de *b*, e assim por diante até a classificação menos prioritária.

Nos estados de fila, os chamados são atendidos de acordo com a prioridade. A Figura 5 mostra como são feitas as transições de estado para um sistema com três classes de prioridades. As transições são feitas da seguinte forma: caso haja um chamado *a* em espera, nenhum outro é atendido antes; caso haja mais de um chamado com a mesma prioridade, estes são atendidos seguindo a disciplina FCFS entre si.

Figura 5 – Transições de estado em um sistema com prioridade em fila.



Fonte: Souza *et al.* (2015).

As transições de estado possibilitam construir as equações de equilíbrio para os estados de fila com prioridade, a Equação (13) mostra o exemplo de equação de equilíbrio para o mesmo estado $\{ab\}$ da Figura anterior.

$$(\lambda + \mu)P_{\{ab\}} = \lambda_a P_{\{b\}} + \lambda_b P_{\{a\}} + \mu P_{\{aab\}} \quad (13)$$

Em que:

- λ_a e λ_b representam as taxas de chegada de chamados com prioridades *a* e *b*, respectivamente.

O número de estados de fila que o sistema possuirá é calculado pela Equação (14). Conforme mencionado anteriormente, o modelo hipercubo considera apenas servidores livres $\{0\}$ ou ocupados $\{1\}$, o modelo terá um total de $2^N + Q$ estados.

$$Q = \binom{r + L}{L} - 1 \quad (14)$$

Em que:

- *r* é o número de classes de chamados; e
- *L* é o número máximo de usuários aceitos na fila.

Nestes modelos pode-se obter medidas específicas para cada tipo de chamado, como frequência de despacho, tempos de espera, tempos de viagem, etc.

Se definirmos n_r o número de usuários da classe r na fila, a probabilidade de haver *j* usuários desta classe na fila, $P(n_r = j)$, é dada pela soma das probabilidades associadas com os estados da fila em que a Equação (15) é obedecida.

$$P(n_r = j) = \sum_{\forall S \; s.t. \; n(r,S)=j} P\{S\} \quad (15)$$

Em que S é o estado da fila e $n(r, S)$ é o número de usuários da classe r em S. Desta distribuição, o número médio de usuários da classe r pode ser determinado pela Equação (16).

$$L_{qr} = \sum_j jP(n_r = j) \quad (16)$$

Pela Fórmula de Little, pode-se calcular o tempo médio de espera para cada prioridade, Equação (17).

$$W_{qr} = L_{qr}/\lambda_r \quad (17)$$

O tempo médio de viagem de um servidor *i* para o subátomo *jr*, é $t_{i,jr}$, dado pela Equação (18) (SOUZA, 2010).

$$t_{i,jr} = \sum_{p=1}^{N_A} \sum_{l \in D} l_{i,jr} \cdot \tau_{pl,jr} \quad (18)$$

Em que:

- $l_{i,jr}$ representa a probabilidade de o servidor i estar localizado no subátomo jr; e
- $\tau_{jr,pl}$ é o tempo de viagem do subátomo pl para o subátomo jr.

O tempo médio de viagem para chamados em fila, $\bar{T}_Q$, é dado pela Equação (19).

$$\bar{T}_Q = \sum_p \sum_l \sum_j \sum_r \frac{\lambda_{pl}\lambda_{jr}}{\lambda^2}\tau_{pl,jr} \tag{19}$$

O tempo médio de viagem ao subátomo *jr* é dado pela Equação (20).

$$\bar{T}_{jr} = \frac{\sum_i f_{i,jr}^{nq} t_{i,jr}}{\sum_i f_{i,jr}^{nq}}(1 - P_S) + \sum_p \sum_l \left(\frac{\lambda_{pl}}{\lambda}\right)\tau_{jr,pl}P_S \tag{20}$$

Em que:

- P_S é a probabilidade de saturação.

O tempo médios de viagem do servidor *i*, $\overline{TU}_i$, é obtido pela Equação (21).

$$\overline{TU}_i = \frac{\sum_j \sum_r f_{i,jr}^{nq} t_{i,jr} + \left(\bar{T}_Q P_S\right)\frac{\mu_i}{\mu}}{\sum_j \sum_r f_{i,jr}^{nq} + P_S\frac{\mu_i}{\mu}} \tag{21}$$

Uma aplicação dessa extensão pode ser encontrada em Souza *et al.* (2015) e Souza (2010). O SAMU de Ribeirão Preto é estudado em diferentes períodos do dia (manhã, tarde e noite) de forma independente. Tem suas taxas de chegada aumentadas e retirada de servidores para a avaliação de cenários alternativos. Assim como avalia o impacto de atender chamados de remoção.

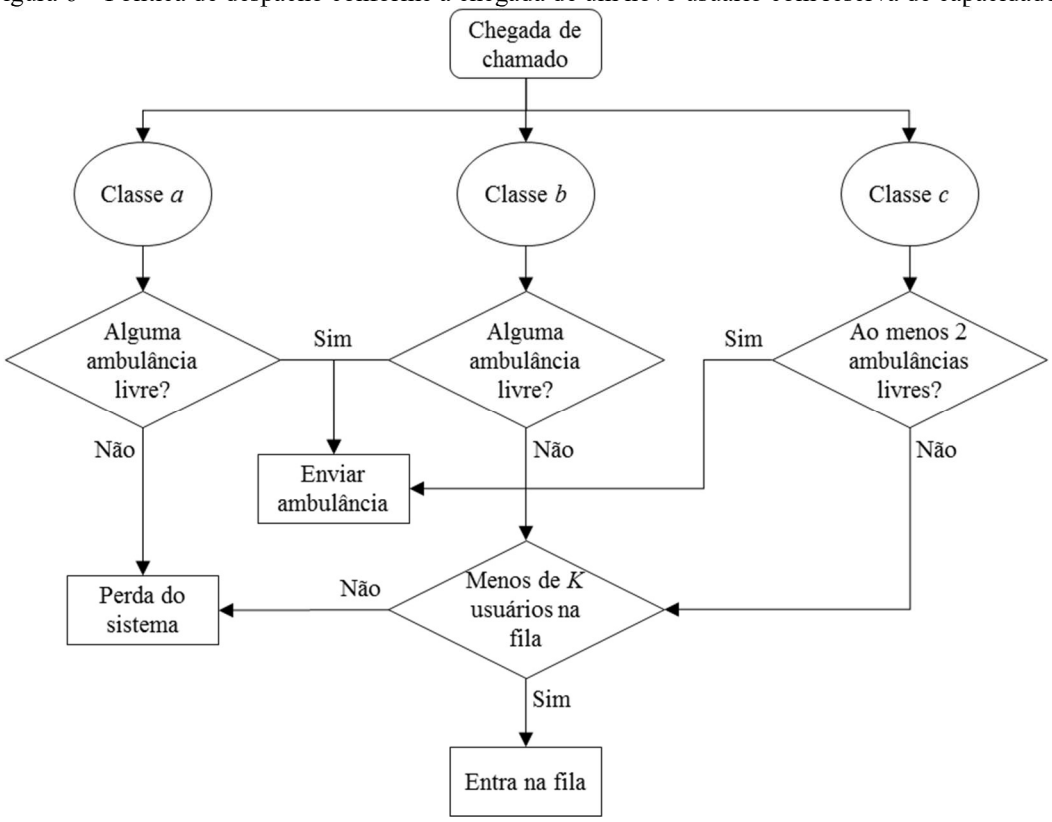
2.6 Extensão 4: Reserva de capacidade

Quando a interrupção de um atendimento não é uma alternativa viável para atender a chamados de maior prioridade, uma alternativa é utilizar reserva de capacidade dos servidores. A reserva de servidores é uma técnica de gestão cuja finalidade é aumentar a probabilidade de um chamado de alta prioridade encontrar um servidor disponível na sua chegada, ela é definida sobre a política de despacho do sistema. Dessa forma, pode ser pertinente enviar para a fila de espera chamados de baixa prioridade, enquanto um determinado número de servidores não estiver disponível (IANNONI *et al.*, 2015).

Iannoni *et al.* (2015) traz um exemplo de reserva de capacidade para um sistema com três classes de usuários, em que cada classe possui um procedimento para realizar o despacho. A Figura 6 traz um esquema que mostra uma possível política de despacho para este tipo de problema. Observe que a prioridade *a* recebe o maior nível de preferência, não sendo permitida

sua entrada em fila caso não haja uma ambulância disponível. Por outro lado, a prioridade *c* recebe a menor prioridade, é interessante que ao menos uma ambulância fique disponível, por isso avalia-se se há ao menos duas disponíveis antes de enviar. É importante ressaltar que desta maneira, pode-se haver formação de fila de espera, mesmo com servidores disponíveis para atendimento.

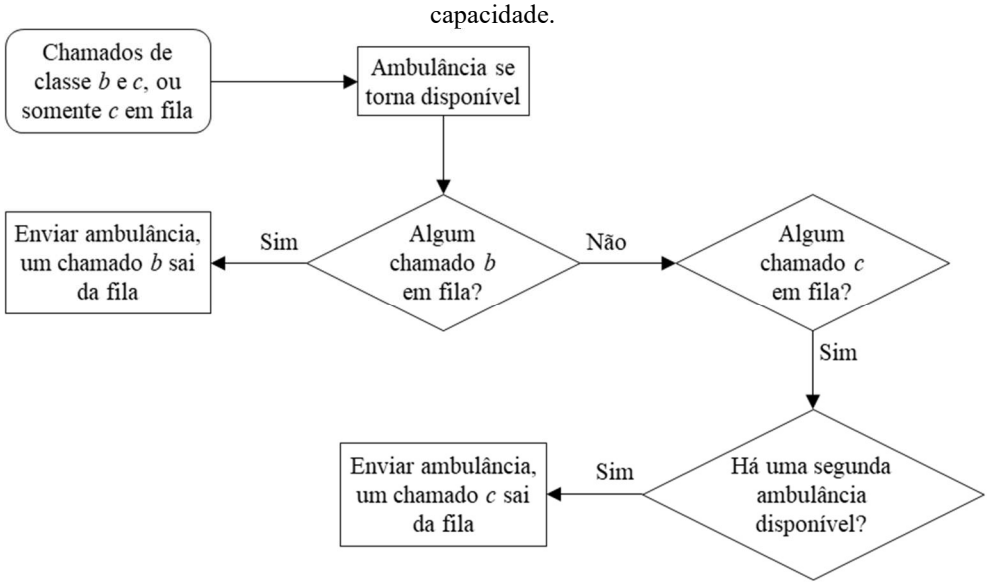
Figura 6 – Política de despacho conforme a chegada de um novo usuário com reserva de capacidade.



Fonte: Iannoni *et al.* (2015, p54) adaptado.

Ainda neste exemplo, pode-se observar que a política de despacho para os chamados também é diferenciada na fila de espera, conforme a Figura 7. Os chamados de prioridade *b* são atendidos antes dos chamados de prioridade *c* e estes apenas são atendidos quando houver ao menos duas ambulâncias livres.

Figura 7 – Política de despacho para os chamados em fila conforme um servidor fica disponível com reserva de capacidade.



Fonte: Iannoni *et al.* (2015, p54) adaptado.

De acordo com Iannoni *et al.* (2015), esta decisão do exemplo de reservar a capacidade de apenas uma ambulância pode ser observada na operação do SAMU de Paris. Contudo, a aplicação do modelo hipercubo com esta extensão foi limitada a um modelo reduzido com três servidores.

Outros exemplos de aplicação de reserva de capacidade podem ser encontrados na literatura sem utilizar o modelo hipercubo. Taylor e Templeton (1980) apresenta dois modelos com reserva de capacidade para aplicações em sistemas médicos emergenciais urbanos com duas classes de usuários. Sacks *et al.* (1993) estendem o modelo para a aplicação no despacho de viaturas policiais, juntamente com uma heurística para obtenção do número ideal de viaturas a serem reservadas. Em um aplicação em *call-centers*, Gurvich *et al.* (2008) utiliza a reserva de capacidade com um procedimento de otimização aproximado.

2.7 Outras extensões

Várias outras extensões do modelo hipercubo podem ser encontradas na literatura. Destacam-se aqui apenas algumas mais relacionadas aos objetivos deste trabalho, conforme levantado durante a revisão de literatura.

2.7.1 Despacho múltiplo

Em uma das hipóteses de aplicação do modelo hipercubo clássico, apenas um servidor pode ser enviado para realizar um atendimento. Esta hipótese, normalmente, não é razoável

para departamentos de polícia cujas unidades possuem apenas um policial, onde duas unidades são enviadas para chamadas potencialmente perigosas. Também não é razoável em serviços médicos emergenciais com ambulâncias para transporte dos usuários e unidades bem equipadas com paramédicos (CHELST; BARLACH, 1981).

Para conseguir modelar este tipo de situação, Chelst e Barlach (1981) propõe uma extensão ao modelo hipercubo com dois tipos de chamados, que cada um segue um processo de Poisson. Os chamados do Tipo 1 são atendidos por apenas um servidor com tempo médio de serviço de $1/\mu$ exponencialmente distribuído. Para os chamados de Tipo 2, dois servidores são enviados e seus tempos médios de serviço são independentes e exponencialmente distribuídos com média $1/\mu$. Como resultado, duas unidades atendendo a um chamado de Tipo 2, podem ser tratadas como duas unidades atendendo a dois chamados de Tipo 1, resultando em um hipercubo 3^N. A principal fraqueza desta extensão é que não se assegura que a primeira unidade a chegar no destino terá um tempo médio de serviço maior, por ter que iniciar o atendimento para todas as vítimas, enquanto a segunda unidade não chegar.

Chelst e Barlach (1981) consideram um sistema que não aceita fila. De forma que, quando houver apenas uma unidade disponível, esta atenderá (sozinha) quaisquer ocorrências que chegarem. Se todas as unidades estiverem ocupadas, os chamados serão considerados perda do sistema.

Esta extensão ainda possibilita calcular outras medidas de desempenho, além das definidas no hipercubo clássico. Elas são voltadas para os chamados de Tipo 2, por exemplo: tempo médio de viagem para os chamados do Tipo 2; tempo médio de viagem para o servidor preferencial para chamados de Tipo 2; tempo médio de viagem para o *backup* de chamados de Tipo 2; tempo médio de viagem para a chegada do primeiro servidor; tempo médio de viagem para a chegada do segundo servidor; a média e a distribuição do intervalo entre as chegadas do primeiro e segundo servidores; e a fração de chamados de Tipo 2 cujo servidor i é o primeiro a chegar ao local da ocorrência.

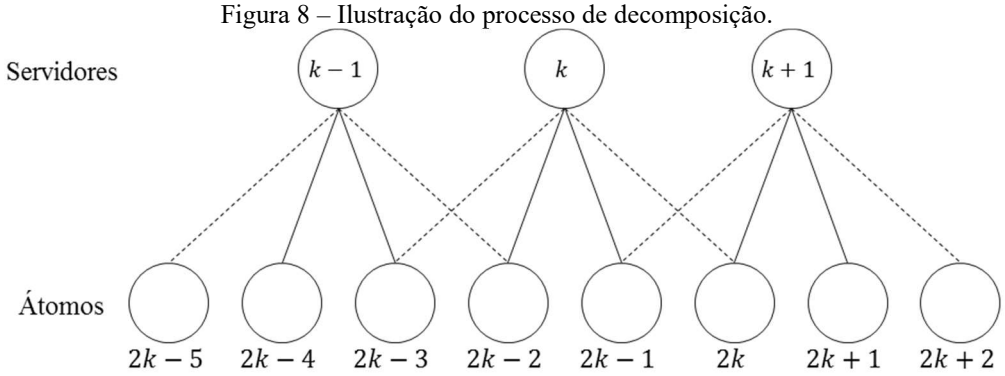
Aplicações desta extensão podem ser vistas em Chelst e Barlach (1981) sobre um sistema urbano de patrulhamento policial em New Haven. Iannoni (2005), Iannoni e Morabito (2006) e Iannoni *et al.* (2008; 2009) utilizam esta extensão, juntamente do backup parcial em um SAE de uma rodovia no Brasil.

Chelst e Barlach (1981) ainda afirma que essa extensão tem potencial de aplicação quando dois sistemas são modelados simultaneamente. Por exemplo, um sistema de bombeiros que, eventualmente, precisa do suporte de uma ambulância para atender aos chamados.

2.7.2 Decomposição do sistema

Atkinson *et al.* (2006) utiliza uma abordagem para modelar um sistema de filas em uma rodovia. O sistema sofre uma decomposição para que sejam resolvidos modelos com três servidores sessão por sessão da rodovia, ele utiliza o hipercubo clássico. A Figura 8 ilustra o processo de decomposição sofrida pelo sistema. As linhas contínuas indicam os átomos em que os servidores são preferenciais. Enquanto as linhas tracejadas indicam os átomos que onde são *backups*.

Figura 8 – Ilustração do processo de decomposição.



Fonte: Atkinson *et al.* (2006, p383) adaptado.

Resolvidos as decomposições de $(k-1, k, k+1)$, $(k, k+1, k+2)$, até o último servidor, encontra-se uma média por meio de um processo iterativo com as intensidades das taxas de chegada para o servidor k. Com isso, é possível encontrar a carga de trabalho para os servidores do sistema.

2.7.3 Agregação de estados no hipercubo 3^N

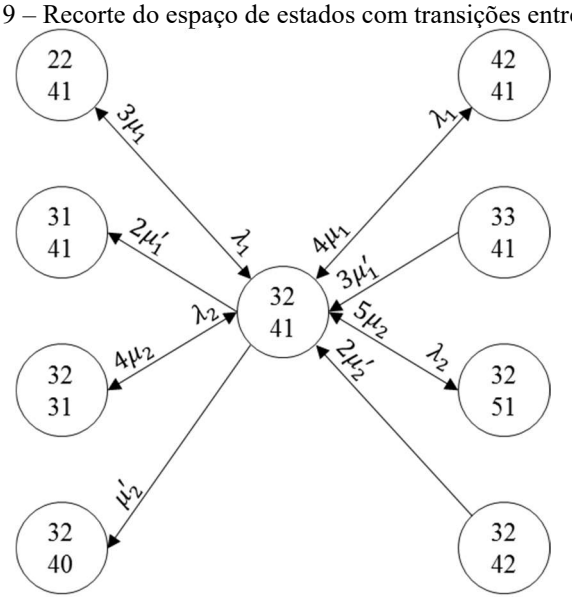
Boyaci e Geroliminis (2015) propõe um modelo conjunto partindo da ideia que os tempos de viagem dos servidores para chamados fora de sua área primária de atendimento não são desprezíveis e devem ser considerados durante a modelagem. A partir disso, são propostos dois modelos para atuarem conjuntamente: um modelo hipercubo 3^N e um modelo agregado.

A abordagem se inicia com o modelo hipercubo com 3 estados possíveis para cada servidor, sendo eles: (0) livre, (1) ocupado em sua área primária e (2) ocupado em área fora da área primária. Como esse modelo tem um alto custo computacional, é proposto que o sistema seja bipartido em sistemas menores até que cada um possua uma quantidade definida de servidores. Essas partições são modeladas utilizando o modelo de 3 estados.

Após a resolução dos modelos hipercubo 3^N para todas partições, é feito o caminho inverso da bipartição, unindo os sistemas utilizando o modelo agregado. O modelo agregado trabalha com dois grupos de servidores, cada grupo pertencente a das partições unidas. Os grupos de servidores são representados por *bins*, dois valores numéricos que indicam o estado do sistema. Por exemplo: o estado {12} representa que há um servidor (qualquer servidor do grupo) ocupado com um chamado em sua área primária (sua partição original), enquanto outros dois (também desconhecidos em meio ao grupo) estão ocupados com chamados fora de sua área primária (partição adjacente a qual seu sistema está se fundindo). O modelo sempre é resolvido com dois *bins*, um de cada partição que se funde. Por exemplo: {12|11} o primeiro *bin* é exatamente igual ao apresentado no exemplo anterior, enquanto o segundo é o que representa a outra partição, em que há um servidor ocupado em sua própria partição e outro ocupado fora dela. Esse modelo é resolvido sucessivamente até que todas partições sejam fundidas no sistema original.

A Figura 9 mostra um recorte de exemplo para um sistema com dois *bins* de seis servidores cada e sua preferência de despacho. O primeiro *bin* é prioritário para atender à região 1 e o segundo *bin* para atender a região 2. Note que cada *bin* possui duas taxas de serviço diferentes, por exemplo, a taxa μ_1 é a taxa de serviço do primeiro *bin* em sua área primária de atendimento, a taxa μ_1' é a taxa de serviço do primeiro *bin* em sua área secundária de atendimento, em que o tempo de viagem é maior e não desprezível.

Figura 9 – Recorte do espaço de estados com transições entre *bins*.



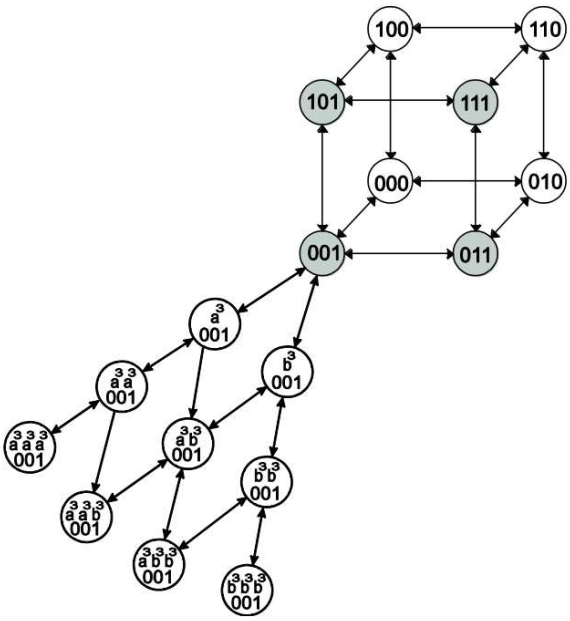
Fonte: Boyaci e Geroliminis (2015, p158) adaptado.

Para que este modelo possua validade é preciso que os servidores atendam uma fração muito pequena de chamados fora de sua área primária, caso contrário, as aproximações feitas durante as fusões surtirão em um viés nos resultados finais do modelo. Além disso, é preciso assegurar o uso de um bom algoritmo de partição do sistema.

2.7.4 *Backup* parcial e prioridade em fila

Rodrigues *et al.* (2017) apresenta uma extensão capaz de trabalhar com filas com prioridade de atendimento em sistemas com restrições geográficas ao atendimento dos chamados. Para tanto, diferencia-se cada estado da fila de acordo com a prioridade recebida e seu átomo de origem, como mostra a Figura 10 com um recorte de um sistema com três servidores. Foi aplicado em um estudo de caso no setor de manutenção e borracharia dentro do setor sucroalcooleiro, onde as distâncias entre os átomos são grandes e o acesso a alguns átomos é inviável dependendo da origem do servidor.

Figura 10 – Recorte com os estado de fila para um sistema com *backup* parcial e prioridade em fila.



Fonte: Rodrigues *et al.* (2017, p96).

3 MODELOS DE FILAS MARKOVIANOS VARIÁVEIS NO TEMPO

Neste capítulo, apresenta-se inicialmente os modelos de programação (*scheduling*) de servidores que utilizam modelos de filas Markovianos $M(t)/M/s(t)$, seguindo aos modelos que utilizam aproximações estacionárias para avaliar sistemas de fila ao longo do dia. Além disso, há uma descrição dos modelos não-estacionários e algumas de suas aplicações utilizando modelos exatos. Para uma revisão completa do estudo de sistemas não-estacionários, considerando processos diferentes de chegada e serviço, recomenda-se a leitura de Schwarz *et al.* (2016) e Defraeye e Nieuwenhuyse (2016) e as referências de ambos.

3.1 Programação (*Scheduling*) de servidores

Programar os servidores é uma atividade inerente ao setor de serviços. Caixas de banco, praças de pedágio, restaurantes, lojas, *call-centers*, são apenas alguns dos exemplos de serviços que necessitam programar seus servidores para atender a uma demanda aleatória e variável no tempo (INGOLFSSON *et al.*, 2010).

Nesse contexto, Buffa *et al.* (1976) apresenta uma abordagem para o processo de programação de servidores em quatro passos:

1. Prever a demanda período a período para sistema, sejam períodos de 15 minutos, 30 minutos, 1 hora, etc.;
2. Converter as previsões período a período em requisitos mínimos de servidores;
3. Determinar o conjunto dos possíveis turnos; e
4. Selecionar o conjunto de turnos que minimize os custos operacionais, enquanto fornece um número mínimo de servidores em cada período.

Atualizações de curto prazo nos programas também podem ser consideradas um passo adicional (DEFRAEYE; NIEUWENHUYSEN, 2016).

O primeiro trabalho sobre o tema é visto em Edie (1954), que realizou um estudo sobre congestionamentos em praças de pedágios e utilizou uma combinação de análises empíricas e modelos de teoria das filas estacionários para gerar os requisitos de servidores para garantir um nível de serviço. Em resposta a este trabalho, Dantzig (1954) mostrou uma maneira de utilizar programação linear inteira para encontrar o programa que atendesse aos requisitos minimizando o número de servidores necessários.

Keith (1979) adaptou o modelo de Dantzig (1954) por considerar o custo computacional de se resolver um problema de programação inteira elevado, a adaptação foi dividida em duas etapas. A primeira etapa é resolver um problema de programação linear de mesma formulação

a Dantzig (1954), porém considerando variáveis reais. O segundo passo é utilizar uma heurística de arredondamento para obter o programa de turnos final. Além disso, o modelo foi estendido para substituir o requisito de número mínimo de servidores por um número desejável de servidores.

Thompson (1997) desenvolveu dois modelos de programação de servidores que podem ser vistos como extensões do modelo de Keith (1979), já que buscam contornar três limitações dele. A primeira limitação é a inabilidade de garantir a entrega de um nível de serviço agregado de forma a minimizar o custo. A segunda limitação é a sua falha em garantir um número mínimo de servidores para fornecer um nível de serviço aceitável em cada período. A terceira limitação levantada é a incapacidade de o modelo maximizar alguma medida de desempenho a partir de um número fixo de servidores. Por isso, o primeiro modelo de Thompson (1997) busca minimizar os custos do sistema, enquanto o segundo busca maximizar o nível de serviço prestado. Ambos modelos evitam as limitações do modelo de Keith (1979) ao adicionar restrições ao nível de serviço considerando variáveis de falta e de excesso assim como coeficientes gerados de modelos de teoria das filas estacionários.

Ao estudar a programação de viaturas policiais, Kolesar *et al.* (1975) estendeu a abordagem de Dantzig (1954) ao realizar a programação de servidores seguindo cinco passos:

1. Encontrar todos os turnos de trabalho possíveis para os servidores;
2. Estimar o requisito mínimo período a período de servidores utilizando modelos de teoria das filas estacionários;
3. Obter um programa ótimo utilizando o modelo de Dantzig (1954);
4. Avaliar o programa obtido utilizando um modelo de teoria das filas não-estacionário (resolvendo um sistema de equações diferenciais); e
5. Se o programa obtido for satisfatório, parar. Caso contrário, modificar o requisito mínimo período a período de servidores e voltar ao passo 3.

É importante ressaltar que Kolesar *et al.* (1975) não traz um algoritmo formal para o passo 5.

Jennings *et al.* (1996) apresenta um método para calcular os efeitos da variação temporal do sistema de forma a permitir a modelagem de horários de sobrecarga (*rush hours*). Para isso, utiliza uma aproximação baseada em modelos não-estacionários com infinitos servidores.

Ingolfsson *et al.* (2002) apresentou um método que utiliza algoritmo genético para gerar programas. A avaliação dos programas gerados é feita por meio de modelos de teoria das filas não-estacionários. Ingolfsson *et al.* (2010) apresenta um modelo heurístico baseado em programação inteira para geração de programas, mas neste caso a avaliação é feita por aproximações ao modelo não-estacionário de teoria das filas.

3.2 Aproximações para sistemas variáveis no tempo

Mesmo que as taxas de chegada de um serviço sejam altamente variáveis ao longo do tempo, pode ser possível utilizar modelos de filas estacionários para levantar requisitos de pessoal, por exemplo. É importante ressaltar que, normalmente, é inadequado utilizar uma taxa média de chegada de um dia inteiro. Por outro lado, é possível utilizar modelos estacionários para estudos de sistemas não-estacionários, separando o tempo (um dia, por exemplo) em intervalos e aplicando modelos estacionários em cada um. Isso costuma ser válido quando os tempos de serviço são muito pequenos e o nível do serviço é elevado (GREEN *et al.*, 2007). Para uma explanação a respeito de sistemas estacionários, veja o Anexo A.

A aproximação estacionária simples encontra a média para todos os parâmetros do sistema ao longo do horizonte de tempo estudado. Essa abordagem pode ser dividida em aproximação da época de pico simples e aproximação da hora de pico simples. Na primeira abordagem, utiliza-se o pico instantâneo dos parâmetros para o período todo. A segunda abordagem separa o dia em intervalos e utiliza o intervalo de pico para os dados de entrada no cálculo da performance (SCHWARZ *et al.*, 2016). Ambas abordagens podem ser vistas em Green *et al.* (1995) em um estudo de sua acuracidade.

A aproximação chamada de estacionário independente período a período (SIPP) por Green *et al.* (2001) funciona a partir de intervalos de tempo menores. Esses intervalos são obtidos a partir das regras de programação de um serviço. Cada intervalo tem a média dos parâmetros calculada e utilizada em um modelo estacionário.

Uma abordagem semelhante é a aproximação estacionária por pontos. Neste caso, os intervalos de tempo são muito pequenos, bastante inferiores ao intervalo obtido a partir das regras para programação de servidores. Essa abordagem pode unir diversos pontos para encontrar os requisitos de pessoal para um intervalo, por isso normalmente fornece requisitos maiores do que o SIPP. Sendo assim, suas semelhanças são que ambos utilizam modelos estacionários em períodos independentes (GREEN *et al.*, 2007).

Quando os tempos de serviço são maiores, torna-se mais difícil trabalhar com os modelos estacionários. Nesses casos é preciso considerar picos de atrasos para os usuários após picos da taxa de chegada. Uma saída é ajustar as abordagens anteriores considerando um atraso entre os dados de entrada e a performance resultante. Uma forma de computar o atraso é considerar a taxa média de chegada no intervalo de tempo anterior, sendo a amplitude do intervalo de tempo é igual ao tempo médio de serviço (GREEN *et al.*, 2007).

Schwarz *et al.* (2016) classifica essas abordagens como estacionárias por partes com períodos independentes. Essas abordagens têm como vantagem a simplicidade computacional. Contudo, erros podem emergir já que o comportamento transiente é negligenciado e os períodos são tratados como independentes.

Outras abordagens foram desenvolvidas a partir de aproximações estacionárias, contudo sem considerar intervalos de tempo independentes. Pode-se citar dois exemplos, o atraso continuado estacionário e a técnica de transformação coordenada. A primeira foi desenvolvida por Stolletz (2008). Essa abordagem também divide o horizonte de tempo em intervalos e aplica fórmulas estacionárias, mas o acúmulo dos chamados não atendidos em um intervalo é carregado para o intervalo seguinte onde é considerado na avaliação do desempenho. A segunda abordagem foi proposta por Kimber *et al.* (1977). Essa abordagem utiliza um modelo parcialmente baseado em uma aproximação determinística de fluídos para fornecer medidas precisas em períodos de sobrecarga.

Schwarz *et al.* (2016) classifica essas abordagens como estacionárias por partes com períodos interligados. Esses métodos apresentam como vantagem a possibilidade de considerar o comportamento transiente da performance de um sistema, além de possibilitarem o estudo de sistemas temporariamente sobrecarregados.

Além das aproximações estacionárias, os modelos com parâmetros constantes em pequenos intervalos de tempo podem ser analisados pelos modelos transientes por partes, conforme classificação de Schwarz *et al.* (2016). Dentre esses modelos, três grandes grupos emergem. O primeiro grupo é formado por abordagens baseadas em modelos transientes, onde a situação final de um período é tratada como a situação inicial do período seguinte. O segundo grande grupo utiliza uma técnica chamada de aproximação de uniformização (ou randomização), essa técnica transforma cadeias de Markov de tempo contínuo em cadeias de Markov de tempo discreto. O terceiro grupo utiliza uma abordagem de tempo discreto, onde se substitui o tempo contínuo por pontos discretos quando o sistema é observado.

Por fim, além das abordagens aqui apresentadas muito brevemente, existem outras baseadas em características modificadas do sistema. Para um estudo detalhado de todas abordagens apresentadas e comparações entre elas, sugere-se a leitura de Schwarz *et al.* (2016) e Green *et al.* (2007) e suas respectivas referências. Nas seções seguintes deste capítulo, é apresentado o modelo que descreve o comportamento exato para sistemas não-estacionários, o qual normalmente serve de *benchmark* para a avaliação das aproximações citadas.

3.3 Modelos não-estacionários para sistemas variáveis no tempo

Em realidade, existem muitos sistemas reais que dificilmente atingem o equilíbrio, porque há mudanças ao longo do tempo em suas taxas de serviço, quantidade de servidores, taxa de chegada, etc. Exemplos incluem as operações em *call-centers*, tráfego em aeroportos, hospitais e patrulhas policiais (SCHUWARZ *et al.*, 2016; GREEN *et al.*, 2001). Há diversos modelos de filas que trabalham com parâmetros variáveis no tempo (*time-dependent*) como as taxas de chegada (MASSEY; WHITT, 1997), taxas de serviço (ROTHKOPF; OREN, 1979), número de servidores (JUNG; LEE, 1989), capacidade do sistema (INGOLFSSON *et al.*, 2002), etc. Devido a tais alterações temporais, as medidas de desempenho de um sistema podem ser substancialmente impactadas, em relação aos modelos em equilíbrio, e, por isso, devem ser levadas em conta na construção dos modelos (SCHUWARZ *et al.*, 2016).

Green *et al.* (2001) mostra que os modelos classificados como SIPP não são uma boa aproximação para sistemas reais. Também traz a consideração que esses modelos aproximados podem gerar resultados acurados quando a frequência de eventos do sistema (taxa de chegada + taxa total de serviço do sistema) é maior que o período natural do sistema, e a taxa de chegada é sempre menor que a taxa total de serviço do sistema. Contudo, para sistemas com uma frequência de eventos pequena ou que passa por um período de *rush*, em que as taxas de chegada são superiores à taxa de serviço total, essa aproximação não é boa.

Seguindo as notações de Kendall (1953), as representações dos sistemas devem mostrar quais parâmetros variam com o tempo. Por exemplo, um sistema $M(t)/M/s(t)$ possui taxa de chegada e número de servidores variáveis no tempo, mas a taxa de serviço individual é constante. Assim, a taxa de chegada é um processo de Poisson heterogêneo (KIM; WHITT, 2014), e o tempo de serviço é exponencialmente distribuído.

Nesse contexto, Green *et al.* (2001) e Ingolfsson *et al.* (2002) desenvolvem em seus trabalhos modelos de filas considerando as equações de Chapman-Kolmogorov (Equação (22)). Elas formam um sistema de equações diferenciais ordinárias de primeira ordem que representam o comportamento não-estacionário do sistema $M(t)/M/s(t)$ (Anexo A). O modelo pode ser resolvido numericamente pelo método de Runge-Kutta (Anexo B) a partir de uma solução inicial (WHITT, 2007).

$$\frac{dP_0(t)}{dt} = -\lambda(t)P_0(t) + \mu P_1(t) \tag{22}$$

$$\frac{dP_n(t)}{dt} = -(\lambda(t) + n\mu)P_n(t) + \lambda(t)P_{n-1}(t) + (n+1)\mu P_{n+1}(t), \qquad 1 \leq n < s(t)$$

$$\frac{dP_n(t)}{dt} = -(\lambda(t) + s(t)\mu)P_n(t) + \lambda(t)P_{n-1}(t) + s(t)\mu P_{n+1}(t), \qquad n \geq s(t)$$

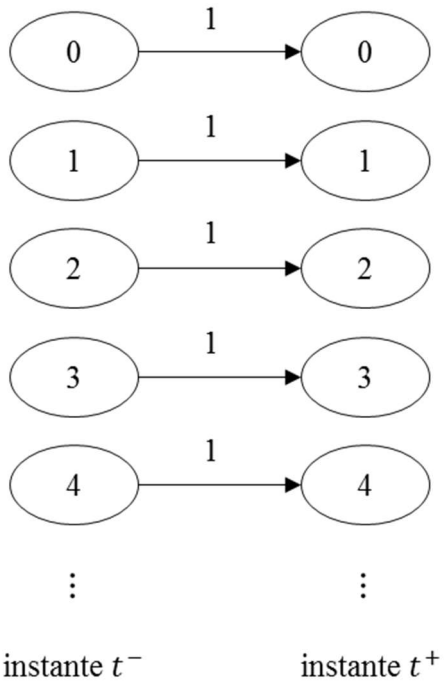
Quando há variações no número de servidores em serviço ao longo do tempo, é importante compreender o comportamento de final de turno desses servidores do sistema. Há duas disciplinas de final de turno para os servidores: preemptivo e exaustivo (INGOLFSSON, 2005). Na disciplina preemptiva, os servidores interrompem o serviço ao final do turno. Ela é mais realista quando os clientes do sistema são objetos. Por exemplo, operações para tapar buracos em ruas após um longo período de chuva, operações de retirada de neve das ruas após uma nevasca (STOLLETZ, 2008). Na disciplina exaustiva, os servidores terminarão o serviço que estiverem realizando antes de saírem ao final do turno. Ela é mais realista quando os usuários são humanos. Por exemplo, operações como médicos em um pronto-socorro, policiais atendendo ocorrências e ambulâncias (FELDMAN *et al.*, 2008).

3.3.1 Disciplina preemptiva

Conforme mencionado anteriormente, na disciplina preemptiva os servidores interrompem o serviço que estiverem realizando no instante do término do turno, momento em que são desligados. Os usuários com o serviço interrompido são realocados para outros servidores ou retornam à fila, caso todos servidores ativos no momento estejam ocupados. Isto significa, em termos de modelagem, que o número de usuários no sistema permanecerá inalterado, no instante t.

Nesse contexto, o vetor de probabilidade no instante logo antes da troca de servidores (instante t^-) e o vetor de probabilidade no instante logo após a troca de servidores (instante t^+) será o mesmo. A Figura 11 ilustra o espaço de estados nesses instantes, onde todos estados têm probabilidade 1 (conforme indicado pelas setas da figura) de permanecerem os mesmos entre t^- e t^+. Essa transição instantânea pode ser escrita conforme uma matriz identidade (INGOLFSSON, 2005).

Figura 11 – Ilustração da transição de estados para a disciplina preemptiva.



O uso da disciplina preemptiva em modelos não-estacionários pode ser notado em vários trabalhos, conforme os exemplos adiante. Embora Margolius (2005) seja o primeiro a revelar o uso explícito da disciplina ao obter equações integrais para a distribuição de probabilidade transiente e o número esperado de usuários em fila em um sistema. Anteriormente, pode-se notar o uso da disciplina preemptiva (muitas vezes não intencional) ao se observar descontinuidades para cima nos gráficos das probabilidades de atraso (ver Seção 3.3). Kolesar *et al.* (1975) utilizou o modelo não-estacionário para verificar o resultado de um problema de programação inteira, baseado no SIPP, na programação de patrulhas policiais de uma delegacia de Nova Iorque. Jennings *et al.* (1996) também utiliza o modelo não-estacionário para avaliar um programa gerado por uma aproximação estacionária em um sistema com demanda variável no tempo. Green *et al.* (2001, 2003) verificam a qualidade de vários métodos aproximados na geração de programas para *call-centers* através do modelo não-estacionário. Ingolfsson *et al.* (2002) propõe o uso do modelo não-estacionário em conjunto com o algoritmo genético na geração de programas que atendam requisitos de minimizar o número total de servidores em serviço, atingir um nível de serviço adequado, etc; e compara com os resultados obtidos utilizando o método apresentado em Kolesar *et al.* (1975). Green e Soares (2007) apresenta uma nota sobre o cálculo exato de algumas medidas de desempenho. Por fim, Ingolfsson *et al.* (2010) traz um problema de programação inteira para atender a requisitos de pessoal em um

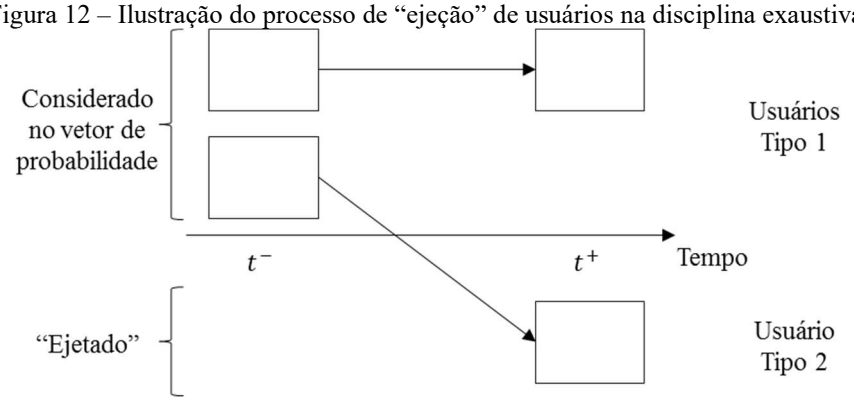
call-center e utiliza um modelo com a disciplina preemptiva para avaliar o programa gerado a partir do nível de serviço.

3.3.2 Disciplina exaustiva

Conforme mencionado anteriormente, na disciplina exaustiva os servidores terminam os atendimentos que estiverem realizando no instante que são desligados, desta maneira os usuários não retornam à fila. Os usuários que estiverem sendo atendidos após o final do turno de seu servidor não afetam as medidas de desempenho do sistema, uma vez que os servidores não ficam disponíveis para novos atendimentos. Por isso, para fins de modelagem, o espaço de estados leva em conta apenas os servidores em seus turnos, no instante t (INGOLFSSON *et al.*, 2007).

Nesse contexto, alguns usuários são "ejetados" do sistema no instante t logo após a mudança de turno, no sentido de que não são mais contabilizados, visto que não impactam as medidas de desempenho dali em diante. Dessa maneira, os usuários precisam ser classificados em dois grupos, usuários do Tipo 1 e Tipo 2. Os usuários do Tipo 1 são aqueles que estão em atendimento por um servidor que está trabalhando em seu turno ou estão em fila. Os usuários do Tipo 2 são aqueles que estão em atendimento por um servidor que está trabalhando além do final de seu turno, terminando um atendimento (INGOLFSSON, 2005). A Figura 12 ilustra o processo pelo qual os usuários do Tipo 2 são "ejetados" do sistema ao final do turno de seu servidor, do instante t^- para o instante t^+. Este evento pode ser considerado uma transição instantânea do sistema.

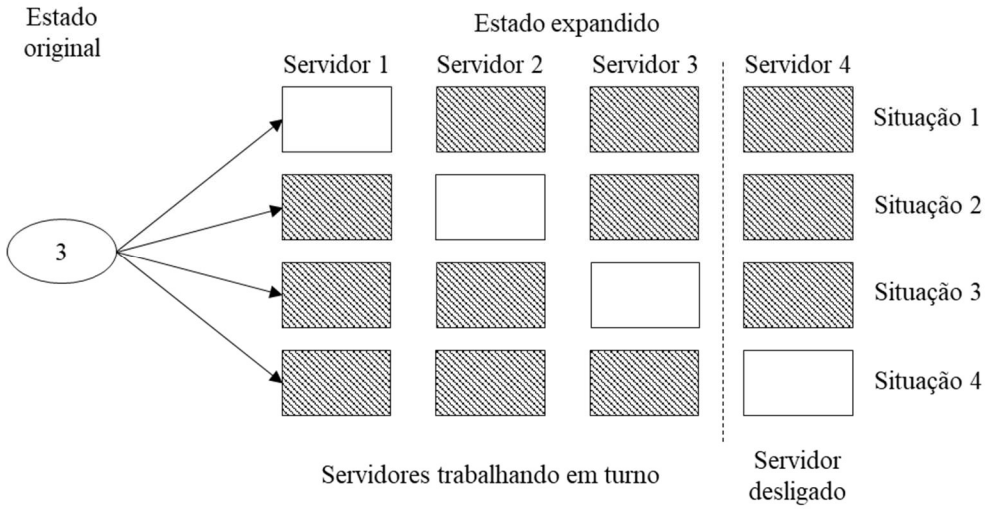
Figura 12 – Ilustração do processo de "ejeção" de usuários na disciplina exaustiva.



Assim, é preciso calcular a probabilidade de um usuário ser "ejetado" para cada estado do sistema. A Figura 13 ilustra o cálculo a partir do espaço de estados expandido (como no modelo hipercubo de Larson (1974)) de forma a observar os servidores individualmente em

todos as situações possíveis. Os retângulos em hachura representam os servidores ocupados, enquanto s retângulos claros indicam o servidor livre. Observe que apenas na situação 4 o servidor em final de turno não está atendendo a nenhum usuário, dessa forma há uma probabilidade de 75% de um usuário ser "ejetado" neste estado de exemplo.

Figura 13 – Esquema para encontrar a probabilidade de um usuário ser "ejetado" na disciplina exaustiva.



A probabilidade de "ejetar" um usuário pode ser calculada a partir de uma distribuição hipergeométrica. O processo corresponde a uma distribuição com uma população igual ao número de servidores logo antes do final do turno, $s(t^-)$, o número de sucessos é igual ao número de servidores saindo, δs. O tamanho da amostra (retirada sem reposição) é igual ao número de usuários no sistema n (INGOLFSSON *et al.*, 2007). A Equação (23) mostra o cálculo da probabilidade de se "ejetar" δn usuários, com os parâmetros δs, $s(t^-)$ e n.

$$\phi(\delta n; \delta s, s(t^-), n) = \frac{\binom{n}{\delta n}\binom{s(t^-)-n}{\delta s-\delta n}}{\binom{s(t^-)}{\delta s}} \tag{23}$$

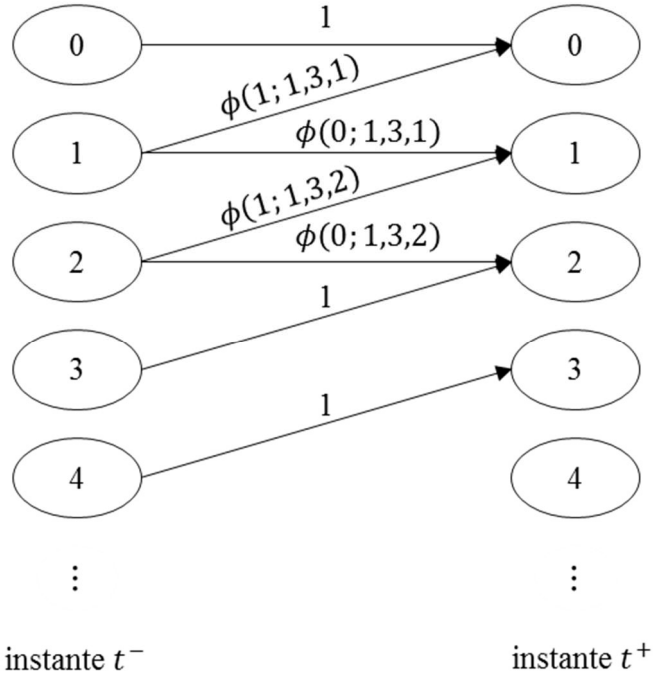
Tendo isso em vista, a transição instantânea de todos estados pode ser representada por uma matriz, $B(t)$, com seus elementos não nulos escritos conforme a Equação (24) (INGOLFSSON *et al.*, 2007).

$$\begin{aligned} b_{n,n-\delta n} &= \phi(\delta n; \delta s, s(t^-), n), \qquad n = 0,1,\dots,s(t^-)-1 \\ &\qquad\qquad \text{e } \left(n-(s(t^-)-\delta s)\right)^+ \le \delta n \le \min(\delta s, n) \\ b_{n,n-\delta s} &= 1, \qquad n = s(t^-), s(t^-)+1, \dots \end{aligned} \tag{24}$$

A Figura 14 ilustra as transições de estado que o sistema passa durante o instante t, mostra um exemplo de um sistema que possui 3 servidores antes do final do período e um deles

será desligado no instante t. Observe que $s(t^-) - \delta s$ não é, necessariamente, o número de servidores trabalhando no próximo período (INGOLFSSON *et al.*, 2007). É importante notar que a probabilidade de todos estados é transmitida integralmente de instante t^- para o instante t^+.

Figura 14 – Ilustração da transição instantânea para um sistema em disciplina exaustiva.



Por fim, a nova solução inicial é obtida pela Equação (25). Onde $\pi(t^+)$ é o vetor de probabilidade dos estados utilizado como solução inicial; $\pi(t^-)$ é o vetor de probabilidade do instante imediatamente anterior à mudança de turno (INGOLFSSON *et al.*, 2007).

$$\pi(t^+) = \pi(t^-)B(t) \tag{25}$$

O uso formal da disciplina exaustiva é mais recente, visto que a solução analítica foi desenvolvida em Ingolfsson (2005). Anterior à solução analítica, Thompson (1993) utilizou aproximações estacionárias considerando a possibilidade de os servidores terminarem algum serviço após o final do turno para determinar requisitos de servidores. Além disso, Feldman *et al.* (2008) ainda não utiliza a disciplina, mas a sugere a importância de ser considerada futuramente, assim como Ingolfsson *et al.* (2010). Ingolfsson *et al.* (2007) verifica a qualidade de diversas aproximações aos modelos não-estacionários considerando um problema com disciplina exaustiva, considerando a randomização como melhor método. Stolletz (2008) utiliza a disciplina exaustiva para avaliar a importância de sua utilização, mesmo em modelos

aproximados. Bassamboo *et al.* (2006) mostra um método de programação de servidores baseado em programação linear.

3.4 Medidas de desempenho

A execução diária de um serviço envolve questões táticas a serem consideradas a fim de melhorar o desempenho do sistema (DIETZ, 2011). O desempenho pode ser avaliado de diferentes maneiras, por meio dos tempos médios de espera em fila, do nível de serviço, entre outras medidas importantes para a operação de um sistema (GREEN; SOARES, 2007). Nesta Seção apresenta-se algumas medidas de desempenho utilizadas para avaliar sistemas de fila não-estacionários, como o nível de serviço, o tempo médio de espera e o número médio de usuários em fila, em um instante t.

3.4.1 Fórmula de Little

Nos modelos não-estacionários a Fórmula de Little ($L = \lambda W$) precisa passar por uma reformulação, uma vez que o tempo médio de permanência no sistema depende de eventos futuros, como a chegada ou saída de um servidor, ou uma mudança no tempo médio de serviço (BERTSIMAS; MOURTZINOU, 1997).

Mantendo o nível de generalidade, Bertsimas e Mourtzinou (1997) reformulou a Fórmula de Little utilizando duas suposições. A primeira suposição é de que o intervalo de tempo entre duas chegadas sucessivas independe das chegadas anteriores, assim como a chegada de novos usuários não afeta o tempo de permanência no sistema dos usuários antigos. A segunda suposição é de que o sistema finaliza o serviço prévio e depois começa a atender aos usuários que chegaram por ordem de chegada após $t = 0$, sendo que o sistema não fica ocioso até terminar os serviços prévios. Tomando essas suposições, a Equação (26) descreve a Fórmula de Little reformulada.

$$E[L(t)] = \int_0^t h(u)P\{S(u) > t - u\}du + \sum_{i=1}^{k} P\{V_i \geq t\} \tag{26}$$

Em que:

- $E[L(t)]$ é a esperança do número de usuários no sistema no instante t;
- $h(u)$ é a taxa de chegada "local" no instante u (seguindo um processo genérico);
- $S(u)$ é o tempo de permanência de um usuário no sistema;

- $P\{S(u) > t - u\}$ é a probabilidade de um usuário permanecer no sistema por um tempo maior do que $t - u$;
- V_i é o tempo até o serviço prévio i ser finalizado; e
- $P\{V_i \geq t\}$ é a probabilidade de o serviço prévio i seja atendido apenas após o instante t;

Observe que diferentemente da Fórmula de Little para sistemas estacionários, $E[L(t)]$ depende completamente da distribuição de $S(t)$, não apenas da sua esperança, e da situação inicial do sistema. Caso assumamos que o sistema se inicie vazio ($\sum_{i=1}^{k} P\{V_i \geq t\} = 0$), que $t \to \infty$ e que os intervalos entre chegadas são sempre positivos e que o intervalo de tempo inicial do sistema é distribuído como o tempo até a chegada de um novo usuário ($h(t) = \lambda$), independendo do tempo, teremos a Fórmula de Little para sistemas estacionários (Equação (27)).

$$E[L(t)] = \lambda \int_0^{\infty} P\{S(u) > t\}dt = \lambda E[S(t)] \tag{27}$$

3.4.2 Nível de serviço

Uma medida importante em sistemas não-estacionários é o nível de serviço, definido como a probabilidade de um novo usuário esperar um tempo menor ou igual a um valor $\tau \geq 0$. Por exemplo, uma meta comum em centrais de atendimento telefônico é a de que 80% dos usuários não esperem mais do que 20 segundos para serem atendidos (INGOLFSSON *et al.*, 2002). O nível de serviço também pode ser entendido como o complementar da probabilidade de atraso. Por exemplo, um nível de serviço de 0,9, ou 90% dos chamados atendidos sem espera é o mesmo que uma probabilidade de atraso de 0,1, ou seja, 10% dos chamados atendidos passaram por espera em fila (GILLARD; KNIGHT, 2014). Dessa maneira, podemos definir a Equação (28) para o nível de serviço.

$$SL(t) = \Pr\{W_Q(t) \leq \tau\} = 1 - \Pr\{W_Q(t) > \tau\} \tag{28}$$

Supondo que as conclusões dos serviços obedecem um processo de Poisson (Anexo A), o nível de serviço pode ser calculado pela Equação (29). Caso o número de servidores sofra um acréscimo de δs no instante $t + \epsilon < t + \tau$, os δs primeiros usuários em fila iniciarão seus atendimentos com certeza antes do instante $t + \tau$ (INGOLFSSON *et al.*, 2007).

$$SL(t) = \begin{cases} 1 - \sum_{i=0}^{K} p_{s(t)+i} \sum_{j=0}^{i} \frac{a^{-j} e^{-a}}{j!}, & s(t+\tau) \leq s(t) \\ 1 - \sum_{i=0}^{K} p_{s(t)+\delta s+i} \sum_{j=0}^{i} \frac{a^{-j} e^{-a}}{j!}, & s(t+\epsilon) = s(t) + \delta s \end{cases} \quad (29)$$

Em que:

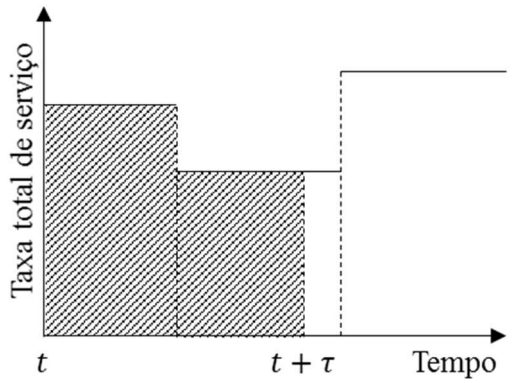
- $\sum_{j=0}^{i} \frac{a^{-j} e^{-a}}{j!}$ é a representação da probabilidade acumulada do estado 0 até i para a distribuição de Poisson, com média de usuários saindo do sistema a, a probabilidade de completar todos os chamados à frente na fila.

Um resultado importante para esta definição é observar o número total de usuários saindo do sistema no intervalo de tempo $(t, t+\tau]$. De forma geral o número total de saídas pode ser dado pela Equação (30), considerando que os tempos de serviço não variam ao longo do tempo (INGOLFSSON *et al.*, 2007).

$$a(t,\tau) \equiv \mu \int_{t}^{t+\tau} s(u)du \quad (30)$$

A Figura 15 ilustra o cálculo da Equação (30) utilizando a área em hachura para obter o número total de saídas no sistema no intervalo $(t, t+\tau]$. Note que, para o cálculo da área, a taxa de serviço total do sistema é seguida. Por exemplo, quando um servidor é desligado dentro do intervalo observado haverá uma redução na taxa de serviço total do sistema, diminuindo o número total de saídas.

Figura 15 – Ilustração do cálculo do número total de saídas para a disciplina preemptiva.



Ingolfsson *et al.* (2007) mostra o cálculo do número total de saídas considerando três cenários básicos. As Equações (31), (32) e (33) mostram estes resultados particulares para a Equação (30). A Equação (31) ilustra o cálculo quando não há alteração no número de servidores entre $(t, t+\tau]$. A Equação (32) mostra o cálculo quando há um acréscimo no

número de servidores no instante $t + \epsilon$, onde $\epsilon < \tau$. Por fim, a Equação (33) mostra o caso de haver redução no número de servidores no sistema. Estas são equações equivalentes à soma da área de retângulos.

$$a(t, \tau) = \mu \tau s(t) \tag{31}$$

$$a(t, \epsilon, \tau) = \mu\big(\epsilon s(t) + (\tau - \epsilon)(s(t) + \delta s)\big) \tag{32}$$

$$a(t, \epsilon, \tau) = \mu\big(\epsilon s(t) + (\tau - \epsilon)(s(t) - \delta s)\big) \tag{33}$$

Estas equações são válidas tanto para a disciplina preemptiva quanto para a exaustiva. Lembrando que, no caso da disciplina exaustiva, o cálculo leva em conta apenas as taxas de serviço para os usuários do Tipo 1.

3.4.3 Tempo médio de espera e número médio de usuários em fila

Além do nível de serviço, outra importante medida de desempenho é o tempo médio de espera em fila para cada instante t. Green e Soares (2007) traz o cálculo para sistemas preemptivos de forma exata, porém a definição do tempo médio de espera em fila pode ser aplicada em sistemas de filas não-preemptivos. O tempo médio de espera em fila do sistema é a esperança dos tempos de espera em fila em cada estado que, como mostra a Equação (34), pode ser simplificada eliminando os estados sem espera em fila.

$$E\left(W_Q(t)\right) = \sum_{i=0}^{K} E\left(W_Q^i(t)\right) p_i(t) = \sum_{i=s(t)}^{K} E\left(W_Q^i(t)\right) p_i(t) \tag{34}$$

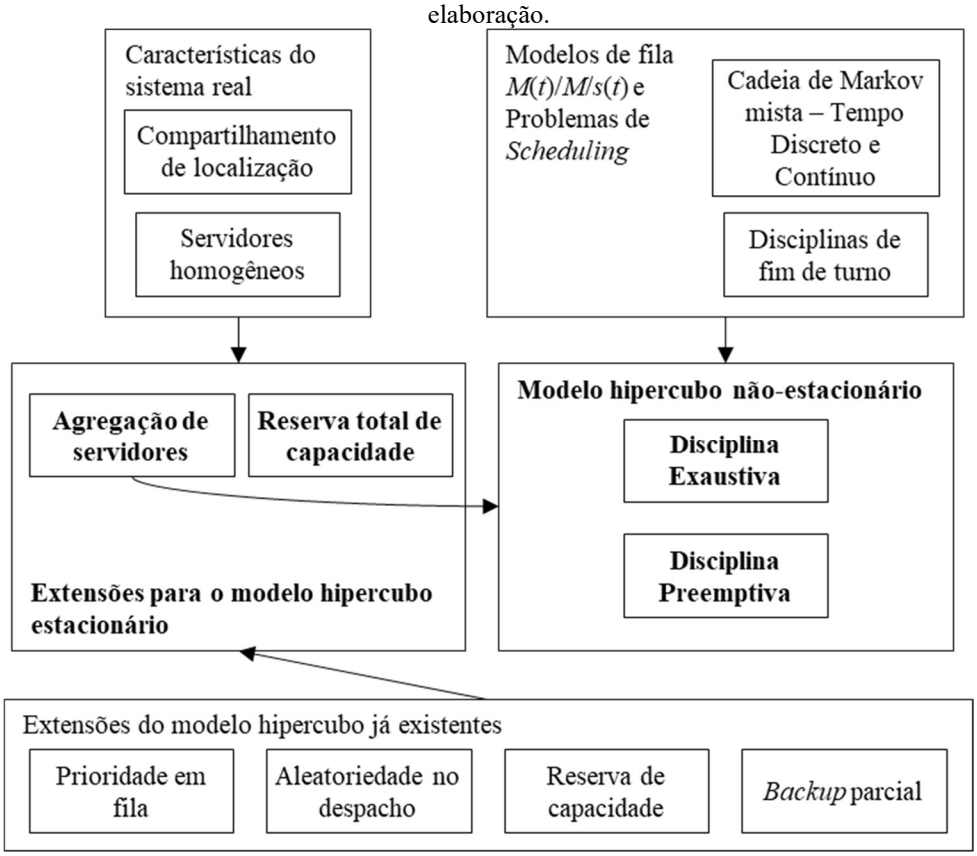
Larson e Odoni (2007) mostram o cálculo do número médio de usuários em fila, conforme a Equação (35). Observa-se que o número médio de usuários em fila depende apenas da situação atual do sistema.

$$L_Q(t) = \sum_{i=s(t)}^{K} \big(i - s(t)\big) p_i(t) \tag{35}$$

4 EXTENSÕES DO MODELO HIPERCUBO PROPOSTAS

Neste capítulo discute-se três extensões do modelo hipercubo de Larson (1974) para o SAMU de Bauru. Primeiramente são apresentadas as extensões para a análise em período de pico do sistema, essas extensões consideram a reserva total de capacidade dos VSA's e a agregação de servidores indistinguíveis. Por fim, é apresentada uma extensão capaz de capturar o comportamento variável no tempo assim como as disciplinas de fim de turno. Todas as extensões são apresentadas por meio de exemplos ilustrativos que já apresentam características encontradas no SAMU-Bauru. A Figura 16 ilustra os modelos desenvolvidos e os conceitos utilizados em sua elaboração.

Figura 16 – Ilustração dos modelos desenvolvidos e os conceitos já presentes na literatura utilizados para a elaboração.



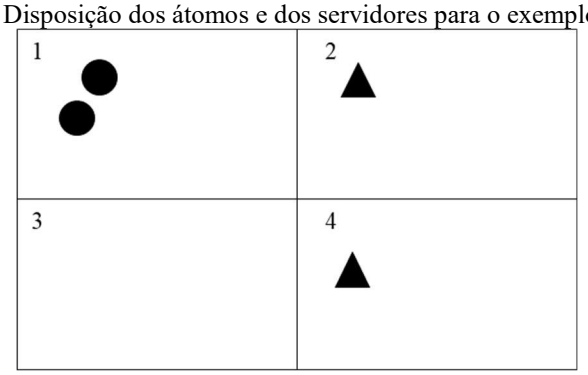
4.1 Extensões do modelo hipercubo para análise estacionária

Esta seção apresenta, por meio de um exemplo ilustrativo as extensões do modelo hipercubo estacionário para o estudo de SAE's em estado estacionário. As extensões representam a agregação de servidores indistinguíveis e a reserva total de capacidade para chamados emergenciais de servidores avançados (como os VSA's).

A agregação de servidores reduz o número de equações necessárias para se representar um sistema de filas, sem que haja perdas com aproximações. A reserva total de capacidade, apresenta a forma como a política de despacho do SAMU-Bauru lida com os chamados, e suas diferentes prioridades na fila e fora dela, em suas am0bulâncias avançadas.

Para apresentar as extensões desenvolvidas, utilizou-se de um exemplo ilustrativo. O exemplo considera um sistema com três átomos e quatro servidores distribuídos conforme a Figura 17, os servidores 1 e 2 no átomo 1 e os servidores 3 e 4 nos átomos 2 e 3 respectivamente.

Figura 17 – Disposição dos átomos e dos servidores para o exemplo ilustrativo.



Lembrando do processo de *layering* (TAKEDA et al., 2007), os chamados dos átomos do exemplo ilustrativo são classificados em duas camadas ou subátomos de acordo com as prioridades *a* e *b* para os chamados de alta e baixa prioridades, respectivamente. A política de despacho do sistema segue uma matriz de preferência de despacho, conforme a Tabela 2. Ao invés de designar um servidor para cada preferência, como a lista de preferência de despacho, a matriz de preferência de despacho designa uma preferência para cada servidor. Assim, pode-se representar, formalmente, situações em que um servidor é escolhido aleatoriamente entre dois ou mais servidores disponíveis com a mesma preferência de despacho. É possível observar que os servidores 1 e 2 possuem a mesma preferência para todos os átomos do sistema, portanto, seu despacho é aleatório em si.

Tabela 2 – Matriz de preferência de despacho para exemplo ilustrativo.

Subátomo	Servidores			
	1	2	3	4
1 *a*	1º	1º	2º	2º
1 *b*	-	-	1º	1º
2 *a*	1º	1º	2º	3º
2 *b*	-	-	1º	2º
3 *a*	1º	1º	2º	2º
3 *b*	-	-	1º	1º

4 *a*	1º	1º	3º	2º
4 *b*	-	-	2º	1º

Os intervalos entre as chegadas do exemplo ilustrativo e os tempos de serviço de todos servidores seguem uma distribuição exponencial. A Tabela 3 mostra os valores das taxas de chegada de usuários em cada subátomo, além de mostrar as taxas de serviço individuais de cada grupo do exemplo ilustrativo.

Tabela 3 – Taxas de chegada e taxas de serviço para o exemplo ilustrativo.

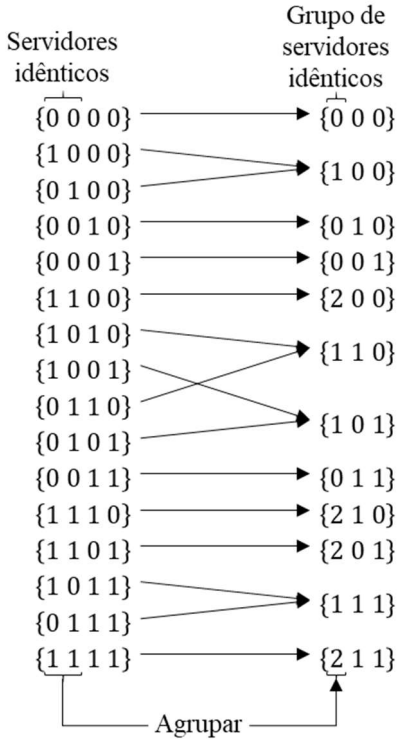
Subátomo	Taxa de chegada	Servidores	Taxa de serviço
1 *a*	0,2	1	1,3
1 *b*	0,4	2	1,3
2 *a*	0,1	3	2,0
2 *b*	0,4	4	1,8
3 *a*	0,1		
3 *b*	0,4		
4 *a*	0,1		
4 *b*	0,3		

4.1.1 Modelo hipercubo com agregação de servidores

Esta extensão tem o propósito de representar o caso em que servidores podem ser agrupados no modelo hipercubo e, assim, diminuir o número de equações de equilíbrio necessárias para resolvê-lo. Luque (2008) sugeriu a possibilidade e algumas condições para que se fizesse a aglutinação de estados no modelo hipercubo a partir de conceitos de aglutinação e decomposição de cadeias de Markov (KIM; SMITH, 1995), mas não desenvolveu técnica ou extensão capaz de tal de forma exata, assim como também não apresentou medidas de desempenho.

Conforme demonstrado, os servidores 1 e 2 possuem a mesma localização, mesma taxa de serviço, mesma preferência de despacho para todos os átomos, são homogêneos. Estas características tornam esses dois servidores completamente indistinguíveis. Dessa maneira, pode-se agrupá-los e diminuir o número de estados do modelo hipercubo. Isto se torna relevante quando se lembra que o número de estados do modelo hipercubo clássico aumenta exponencialmente. A formação de grupos de servidores para o exemplo ilustrativo é feita conforme mostra a Figura 18.

Figura 18 – Estados com servidores não agrupados e servidores agrupados para o exemplo ilustrativo.



Quando é feito o agrupamento de servidores, ocorre uma mudança no cálculo do número de estados do sistema, conforme a Equação (36). Note que quanto mais servidores puderem ser agregados, maior a redução no número de estados.

$$|M| = \prod_{k=1}^{m}(n_k + 1) + Q \tag{36}$$

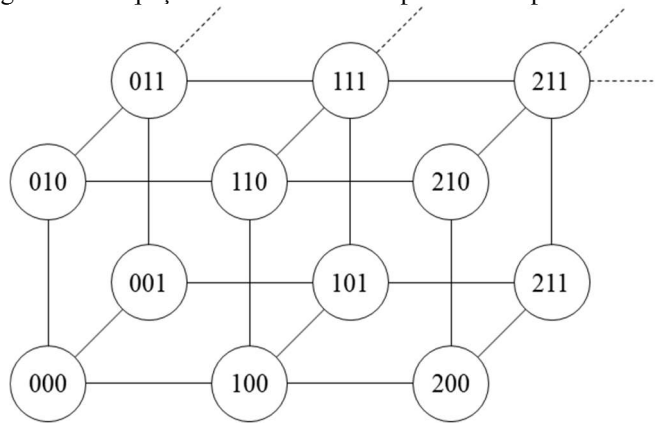
Em que:

- m é o número de grupos;
- n_k é o número de servidores agrupados no grupo k (não é obrigatório os grupos conterem mais do que um servidor).

A Figura 19 ilustra o espaço de estados para o exemplo ilustrativo. São três grupos formados, por isso, três dimensões. É importante ressaltar que, agora os estados do sistema não estão obrigatoriamente nos vértices do cubo (ou hipercubo), portanto, conceitualmente, o modelo agregado deixa de ser um modelo hipercubo. Contudo, devido às semelhanças entre ambos, optou-se por continuar chamando-o como hipercubo. O primeiro grupo é composto pelos servidores 1 e 2, o segundo e terceiro grupos são formados por um servidor cada, os servidores 3 e 4, respectivamente. Assim o número de estados, sem considerar os estados de fila, pode ser calculado pela Equação (36) como $|M| = (2 + 1) \cdot (1 + 1) \cdot (1 + 1) = 12$

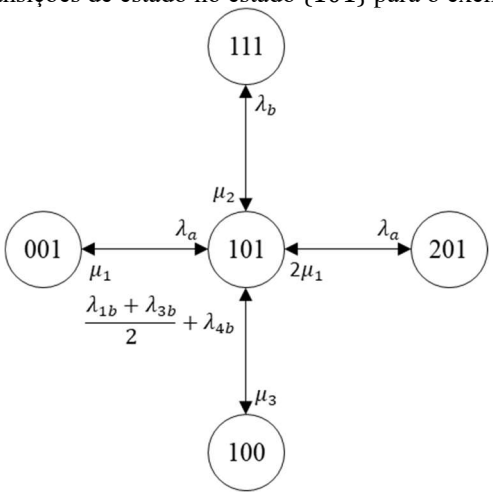
estados. Para se referir aos sistemas formados por agrupamentos, utiliza-se a notação (2,1,1), por exemplo, que indica o número de servidores por grupo.

Figura 19 – Espaço de estados sem fila para o exemplo ilustrativo.



Ao realizar um recorte no espaço de estados, é possível mostrar as transições com mais detalhes. A Figura 20 traz um recorte com foco no estado $\{101\}$ e seus estados adjacentes $\{001\}$, $\{100\}$, $\{210\}$ e $\{111\}$. As transições de estados e a equação de equilíbrio para o estado $\{101\}$ é dada adiante na Equação (37).

Figura 20 – Transições de estado no estado {101} para o exemplo ilustrativo.



As equações de equilíbrio são construídas a partir do conceito em que a taxa de entrada em um estado é igual à taxa de saída do estado. É possível observar que no estado $\{100\}$, metade dos chamados, do tipo *b*, dos átomos 1 e 3, devido à aleatoriedade no despacho, ou qualquer chamado do tipo *b* do átomo 4 levam ao estado $\{101\}$, totalizando uma taxa de $\left(\frac{\lambda_{1b}+\lambda_{3b}}{2}+\lambda_{4b}\right)$. Além disso, a finalização de algum atendimento do grupo 1 no estado $\{201\}$ levam ao estado

$\{101\}$ com taxa $2\mu_1$. Dessa maneira, a equação de equilíbrio do estado $\{101\}$ é escrita conforme a Equação (37).

$$(\lambda + \mu_1 + \mu_3)P_{\{101\}} = \lambda_a P_{\{001\}} + \left(\frac{\lambda_{1b} + \lambda_{3b}}{2} + \lambda_{4b}\right) P_{\{100\}} + \mu_2 P_{\{111\}} + 2\mu_1 P_{\{201\}} \quad (37)$$

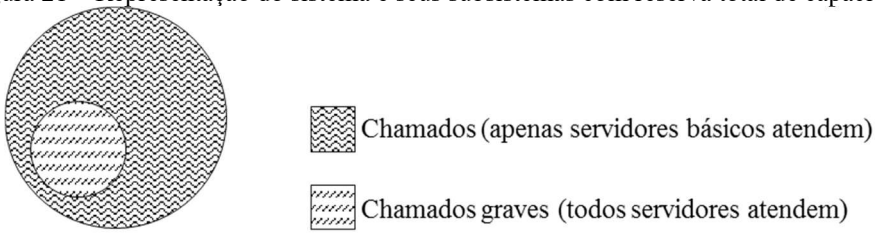
Com a agregação de servidores, foram necessários 12 estados para representar o exemplo, caso não houve agregação seriam necessários 16 estados. Isso resulta em uma redução de 25% com a formação de apenas um grupo de dois servidores. Quanto mais servidores puderem ser agrupados, maior será a redução de estados e a consequente rapidez na resolução computacional.

4.1.2 Modelo hipercubo com reserva total de capacidade

Esta extensão tem o propósito de representar o caso em que os servidores avançados, são reservados para atender apenas aos chamados mais graves de um SAE urbano. Enquanto isso, os servidores básicos ficam disponíveis para atender a todos os chamados, mas preferenciais para os menos graves.

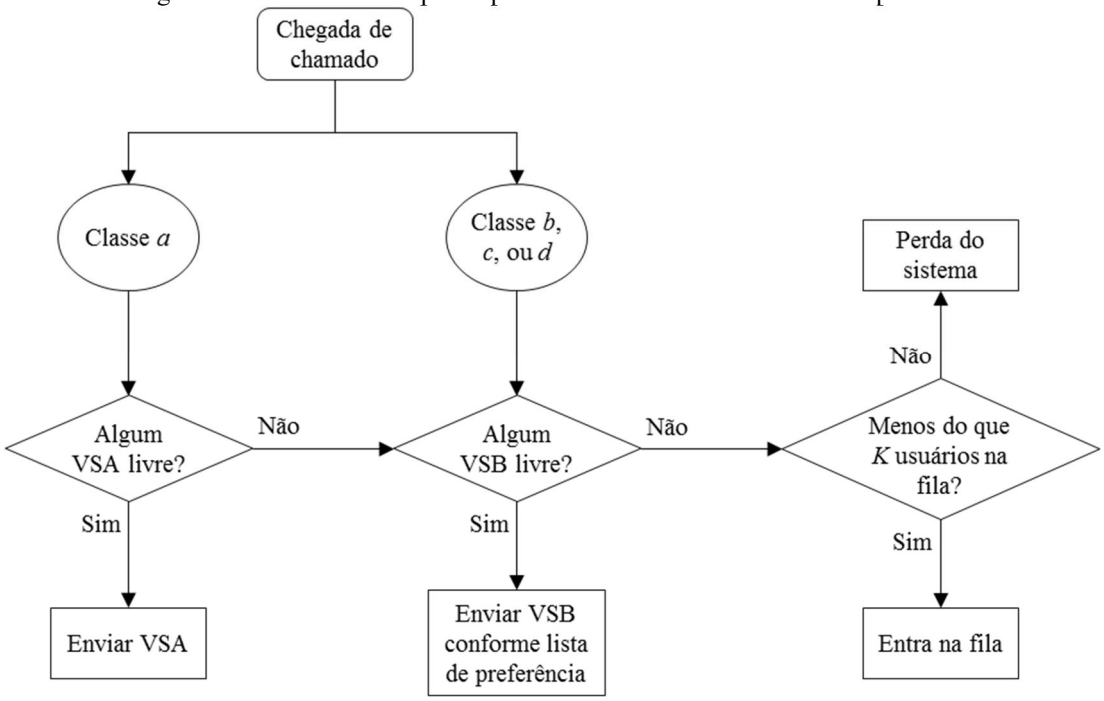
Embora VSA's possuam restrições ao atendimento de certos tipos de chamados a limitação não é física, diferentemente do *backup* parcial. Há uma restrição ao atendimento de determinados subátomos do sistema. Para ilustrar esta situação, pode-se entender que o sistema é composto por dois subsistemas: um menor e atendido por todos os servidores (chamados graves), e outro maior atendido apenas pelos servidores básicos; como mostra a Figura 21.

Figura 21 – Representação do sistema e seus subsistemas com reserva total de capacidade.



A Figura 22 mostra a política de despacho utilizada para modelar o SAMU/Bauru. Os VSA's são preferenciais para o atendimento dos chamados de classe *a*, mas não são considerados no atendimento das outras classes de usuários, aqui chamadas de classes *b*, *c* e *d*. Para o modelo reduzido, as classes *b*, *c* e *d* podem ser entendidas apenas como classe *b*.

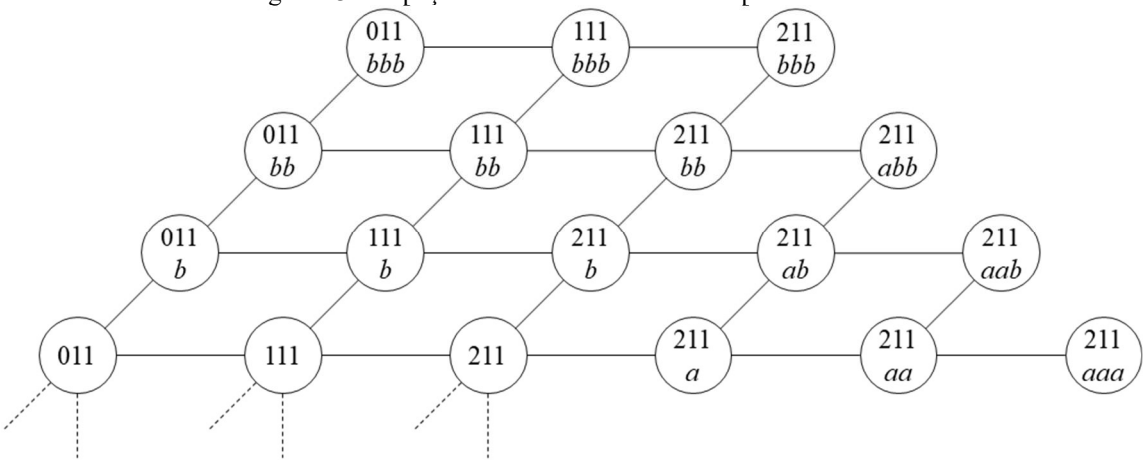
Figura 22 – Política de despacho para sistema com reserva total de capacidade



A princípio, a reserva dos servidores avançados pode sobrecarregar os servidores básicos. Veja que a ideia desse tipo de operação é garantir atendimento adequado aos usuários graves (IANNONI *et al.*, 2015), mantendo a disponibilidade dos servidores avançados. Caso um servidor básico seja enviado no lugar de um avançado, o usuário receberá o atendimento adequado apenas quando chegar a um hospital, aumentando o tempo de resposta percebido pelo usuário e aumentando o risco de vida.

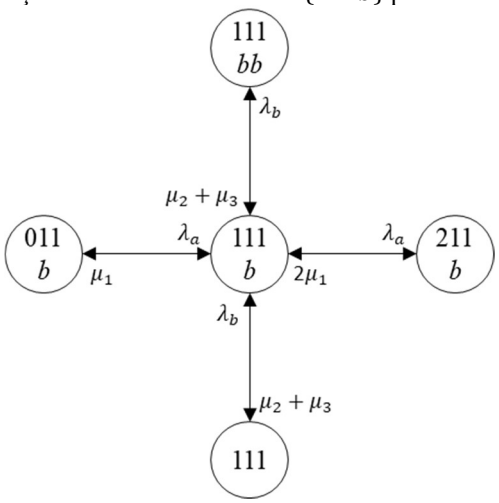
A Figura 23 traz os estados com certa saturação e os estados de fila para o exemplo. Semelhantemente ao modelo de Rodrigues *et al.* (2017), pode haver formação de fila de espera em estados com servidores livres, como o {011}, num caso de semi-saturação (IANNONI *et al.*, 2009). Os demais estados de fila são formados com o sistema completamente saturado, de forma análoga ao modelo de Souza *et al.* (2015). O modelo reduzido, considera um limite de três chamados em espera. É importante ressaltar que, embora haja uma semelhança no espaço de estados com o modelo aproximado de Rodrigues *et al.* (2017), este modelo não é uma aproximação. Portanto, pode ser considerado um modelo Markoviano, já que a origem dos chamados é sempre conhecida e não depende do cálculo de probabilidades condicionais de o sistema estar em cada estado a cada transição.

Figura 23 – Espaço de estados de fila do exemplo ilustrativo.



Um outro recorte, mais específico pode ser feito para entender os fluxos nos estados de fila. A Figura 23 mostra um recorte com foco no estado $\{111b\}$. Este estado ilustra a reserva de capacidade do servidor 1, enquanto os estados completamente saturados ilustram a prioridade em fila.

Figura 24 – Transições de estados no vértice $\{111b\}$ para o exemplo ilustrativo.



Novamente, a equação de equilíbrio é construída considerando que a taxa de saída do estado é igual à taxa de entrada no mesmo estado. É possível observar que apenas no caso de os servidores dos grupos 2 ou 3 finalizarem um atendimento (com taxa $\mu_2 + \mu_3$) que os chamados do tipo b sai da fila, já que o grupo 1 está totalmente reservado para atender aos chamados da classe a. Por outro lado, apenas a chegada de um novo chamado de classe a, pode ocupar o segundo servidor do grupo 1, levando ao estado $\{211b\}$. A equação de equilíbrio que descreve as transições no estado $\{111b\}$ é representada na Equação (38).

$$(\lambda + \mu_1 + \mu_2 + \mu_3)P_{\{111b\}} = \lambda_b P_{\{111\}} + (\mu_2 + \mu_3)P_{\{111bb\}} + \mu_1 P_{\{211b\}} \quad (38)$$

As equações de equilíbrio para cada um dos estados do exemplo ilustrativo são representadas na Equação (39). Observe que as equações dos estados $\{101\}$, $\{011b\}$ e $\{111ab\}$, ilustradas nas Equações (37) e (38).

$$
\begin{aligned}
000 \quad & 2{,}0P_{\{000\}} = 1{,}3P_{\{100\}} + 2{,}0P_{\{010\}} + 1{,}8P_{\{001\}} \\
100 \quad & 3{,}3P_{\{100\}} = 0{,}5P_{\{000\}} + 2P_{\{110\}} + 1{,}8P_{\{101\}} \\
010 \quad & 4{,}0P_{\{010\}} = 0{,}8P_{\{000\}} + 1{,}3P_{\{110\}} + 1{,}8P_{\{011\}} \\
001 \quad & 3{,}8P_{\{001\}} = 0{,}7P_{\{000\}} + 1{,}3P_{\{101\}} + 2{,}0P_{\{011\}} \\
200 \quad & 4{,}6P_{\{200\}} = 0{,}5P_{\{100\}} + 2{,}0P_{\{210\}} + 1{,}8P_{\{201\}} \\
110 \quad & 5{,}3P_{\{110\}} = 0{,}5P_{\{010\}} + 0{,}8P_{\{100\}} + 1{,}8P_{\{111\}} \\
101 \quad & 5{,}1P_{\{101\}} = 0{,}5P_{\{001\}} + 0{,}7P_{\{100\}} + 2{,}0P_{\{111\}} \\
011 \quad & 5{,}8P_{\{011\}} = 1{,}5\left(P_{\{001\}} + P_{\{010\}}\right) + 1{,}3P_{\{111\}} + 3{,}8P_{\{011b\}} \\
210 \quad & 6{,}6P_{\{210\}} = 1{,}05P_{\{200\}} + 0{,}5P_{\{110\}} + 1{,}8P_{\{211\}} \\
201 \quad & 6{,}4P_{\{201\}} = 0{,}95P_{\{200\}} + 0{,}5P_{\{101\}} + 2{,}0P_{\{211\}} \\
111 \quad & 7{,}1P_{\{111\}} = 0{,}5P_{\{011\}} + 1{,}5\left(P_{\{110\}} + P_{\{101\}}\right) + 2{,}6P_{\{211\}} + 3{,}8P_{\{111b\}} \\
211 \quad & 8{,}4P_{\{211\}} = 0{,}5P_{\{111\}} + 2{,}0\left(P_{\{210\}} + P_{\{201\}}\right) + 6{,}4P_{\{211a\}} + 3{,}8P_{\{211b\}} \\
011b \quad & 5{,}8P_{\{011b\}} = 1{,}5P_{\{011\}} + 3{,}8P_{\{011bb\}} + 1{,}3P_{\{111b\}} \\
011bb \quad & 5{,}8P_{\{011bb\}} = 1{,}5P_{\{011b\}} + 3{,}8P_{\{011bbb\}} + 1{,}3P_{\{111bb\}} \\
011bbb \quad & 4{,}3P_{\{011bbb\}} = 1{,}5P_{\{011bb\}} + 1{,}3P_{\{111bb\}} \\
111b \quad & 7{,}1P_{\{111b\}} = 1{,}5P_{\{111\}} + 0{,}5P_{\{011b\}} + 3{,}8P_{\{111bb\}} + 2{,}6P_{\{211b\}} \\
111bb \quad & 7{,}1P_{\{111bb\}} = 1{,}5P_{\{111b\}} + 0{,}5P_{\{011bb\}} + 3{,}8P_{\{111bbb\}} + 2{,}6P_{\{211bb\}} \\
111bbb \quad & 5{,}6P_{\{111bbb\}} = 0{,}5P_{\{011bbb\}} + 1{,}5P_{\{111bb\}} + 2{,}6P_{\{211bbb\}} \\
211a \quad & 8{,}4P_{\{211a\}} = 0{,}5P_{\{211\}} + 6{,}4P_{\{211aa\}} \\
211b \quad & 8{,}4P_{\{211b\}} = 1{,}5P_{\{211\}} + 0{,}5P_{\{111b\}} + 6{,}4P_{\{211ab\}} + 3{,}8P_{\{211bb\}} \\
211aa \quad & 8{,}4P_{\{211aa\}} = 0{,}5P_{\{211a\}} + 6{,}4P_{\{211aaa\}} \\
211ab \quad & 8{,}4P_{\{211ab\}} = 0{,}5P_{\{211b\}} + 1{,}5P_{\{211a\}} + 6{,}4P_{\{211aab\}} \\
211bb \quad & 8{,}4P_{\{211bb\}} = 1{,}5P_{\{211b\}} + 0{,}5P_{\{111bb\}} + 6{,}4P_{\{211abb\}} + 3{,}8P_{\{211bbb\}} \\
211aaa \quad & 6{,}4P_{\{211aaa\}} = 0{,}5P_{\{211aa\}} \\
211aab \quad & 6{,}4P_{\{211aab\}} = 1{,}5P_{\{211aa\}} + 0{,}5P_{\{211ab\}} \\
211abb \quad & 6{,}4P_{\{211abb\}} = 1{,}5P_{\{211ab\}} + 0{,}5P_{\{211bb\}} \\
211bbb \quad & 6{,}4P_{\{211bbb\}} = 1{,}5P_{\{211bb\}} + 0{,}5P_{\{111bbb\}}
\end{aligned}
\quad (39)
$$

4.1.3 Medidas de desempenho para as extensões do modelo estacionário

Para o cálculo da carga de trabalho, é importante lembrar que não se saberá qual servidor de um grupo estará ocupado (a menos que $n_k = 1$). Como a escolha entre os servidores do grupo é aleatória, assume-se que a carga é dividida entre os servidores do grupo. Por exemplo, para o grupo 1 no estado $\{111\}$, tem-se que os servidores 1 e 2 ficaram ocupados metade das vezes, neste estado, cada. A Equação (40) mostra como é feito o cálculo da carga de trabalho ρ_{ki} do servidor ki.

$$\rho_{ki} = \sum_{B \in M} \frac{n_{kB} \cdot P_B}{n_k} \tag{40}$$

Em que:

- B é um estado pertencente ao conjunto de estados M;
- n_{kB} é o número de servidores do grupo k que se encontram ocupados no estado B;
- P_B é a probabilidade do estado B;

A Tabela 4 mostra as cargas de trabalho dos servidores do exemplo ilustrativo. Os servidores pertencentes ao mesmo grupo, obtém a mesma carga de trabalho, como os servidores 1 e 2 do grupo 1. Verificou-se que a agregação de servidores traz os mesmos resultados que o modelo sem agregação de servidores, porém utiliza menos equações de equilíbrio para tal. Além disso, o modelo é Markoviano.

Tabela 4 – Cargas de trabalho para os servidores do exemplo ilustrativo de agregação de servidores.

Grupo	Servidor	Carga de trabalho
1	1	0,2350
	2	0,2350
2	3	0,2035
3	4	0,1933

Por outro lado, as frequências de despacho precisam ser revistas. A Equação (41) mostra o cálculo das frequências de despacho tanto para os chamados sem espera, quanto para os chamados com espera em fila.

$$\begin{aligned} f_{i,j} &= f_{i,jl}^{(nq)} + f_{i,jl}^{(q)} \\ f_{ki,jl}^{(nq)} &= \frac{\lambda_{jl}}{\lambda} \frac{1}{n_k} \sum_{D \in E_{k,jl}} \frac{P_D}{n_{D,jl}} \end{aligned} \tag{41}$$

$$f_{ki,jl}^{(q)} = \frac{\lambda_{jl}}{\lambda} \sum_{E} \left(\frac{\sum_{S \epsilon T_E} \lambda_S}{\lambda} \cdot P_E \right) \frac{\mu_{ki}}{\sum_{A \epsilon M_j} \mu_A}$$

Em que:

- $\sum_E \left(\frac{\sum_{S \epsilon T_E} \lambda_S}{\lambda} \cdot P_E \right)$ é a probabilidade de saturação do sistema;
- $\sum_{S \epsilon T_E} \lambda_S$ é a taxa total de chegada que resultará em espera no estado E, com probabilidade P_E (IANNONI *et al.*, 2009);
- $\sum_{A \epsilon M jl} \mu_A$ taxa de serviço total disponível para atender ao subátomo jl;

Os tempos médios de viagem e os tempos médios de espera não possuem diferenças em seu cálculo entre a extensão apresentada aqui e a extensão sobre prioridade na fila, apresentados em Souza *et al.* (2015). Outras medidas de desempenho importantes podem ser vistas também em Iannoni *et al.* (2015).

4.1.4 Generalização das equações de equilíbrio

Para representar os estados do sistema com agregação de servidores são utilizados vetores compostos por números. A Figura 25 traz um pseudocódigo para a obtenção dos estados do sistema. Os comentários são iniciados pelo símbolo "%".

Figura 25 – Pseudocódigo para geração dos estados do sistema com agregação de servidores sem os estados de fila.

```
% Procedimento para gerar os estados sem fila.

Gerar todos os números (em base binária) de 0 até 2ᴺ-1.
Separar os algarismos dos números em colunas

% Cada coluna deve ser formatada como "double".

% Assim, caso o grupo 1 possua três servidores, as três
% primeiras colunas dos números binários devem
% representar esses três servidores.

Somar as colunas dos servidores agrupados

% Seguindo o exemplo acima, somar as colunas 1, 2 e 3.

Retirar as linhas duplicadas

% Esse procedimento é idêntico ao visto na Figura 18.
```

Para representar os estados do sistema com fila são utilizados vetores compostos pode números também, por conveniência computacional. Os chamados de prioridades *a*, *b*, *c*, ... são

representados pelos números 1, 2, 3, ... sem perda de generalidade. A Figura 26 mostra o pseudocódigo seguido para obtenção dos estados. Um código alternativo pode ser visto em Souza *et al.* (2015), porém sem considerar a reserva de capacidade. Os comentários são iniciados pelo símbolo "%".

Figura 26 – Pseudocódigo para geração dos estados de fila.

```
% Procedimento para gerar os estados de fila.

Verificar o número de prioridades
Verificar quais prioridades têm restrição de atendimento
Verificar estados saturados e semi-saturados

% Gerando estados de fila para o sistema saturado

Para 1 usuário em fila até o limite de capacidade da fila ("for
Q=1:K")
     encontrar todos arranjos com repetição possíveis para os
     chamados em fila;
     ordenar os chamados dos arranjos por prioridade;
     Excluir os arranjos ordenados repetidos;
     Concatenar abaixo dos chamados de fila anteriores;
     % para que as matrizes tenham o mesmo comprimento, preencher
     % os espaços de Q+1 até K com zeros: fila=[Q=1;Q=2;Q=3;...].

Concatenar à direita com estado saturado
% o estado saturado precisa ser repetido para cada linha (estado
% de fila possível) [Estado saturado, fila].

% Como há estados semi-saturados, basta repetir o processo para
% estes sem considerar as prioridades reservadas. Por exemplo,
% considerando prioridades a, b e c, caso os VSA's atendam apenas
% aos chamados a, gerar estados de fila considerando apenas b e
% c.

Repetir o processo para os estados semi-saturados

% Agora o sistema possui uma matriz com os estados do sistema sem
% fila e os estados de fila, como a matriz com os estados de fila
% é mais larga em K unidades, basta:

Preencher a matriz dos estados sem fila K colunas de zeros à
direita
Concatenar os estados com fila abaixo dos demais.
```

A generalização das equações de equilíbrio pode ser vista na Equação (42).

$$P_r\left[\mathbb{I}(\mathbb{L}(r)=1)\sum_j\sum_l \mathbb{I}(\mathbb{R}(r,jl)\neq\emptyset)\lambda_{jl}+\mathbb{I}(\mathbb{L}(r)=0)\lambda+\sum_k\frac{\mathbb{S}(r,k)\mu_k}{n_k}\right]$$
$$=\sum_{q,k:\mathbb{D}(q,r,k)=1}P_q\frac{\mathbb{S}(q,k)\mu_k}{n_k}+\sum_{q,l:\mathbb{Q}(q,r,l)=1}P_q\sum_k\frac{\mathbb{S}(q,k)\mu_k}{n_k} \quad (42)$$
$$+\sum_{q,k:\mathbb{D}(r,q,k)=1}P_q\sum_{jl\in\mathbb{A}(q,k)}\lambda_{jl}+P_q\sum_{q,l:\mathbb{Q}(r,q,l)=1}\lambda_{jl}$$

Em que:

- r e q são estados do sistema, são pertencentes ao conjunto de estados M;
- k indica o grupo de servidores observado;
- jl indica o subátomo observado (átomo j, prioridade l);
- $\mathbb{I}(*)$ é uma função indicador. É igual a 1, se $*$ for verdadeiro e 0, caso contrário;
- $\mathbb{L}(r)$ indica se o estado r possui algum tipo de perda. 0 se não houver nenhuma possibilidade de perda e 1 caso contrário;
- $\mathbb{R}(r,jl)$ indica o os servidores ainda disponíveis para atender ao átomo jl no estado r;
- $\mathbb{S}(r,k)$ mostra o número de servidores do grupo k ocupados no estado r;
- $\mathbb{D}(q,r,k)$ é definida pela Equação (43):

$$\mathbb{D}(q,r,k)=\begin{cases}1, & \text{caso } \mathbb{H}(q,r)=1, \mathbb{S}(q,k)=\mathbb{S}(r,k)+1\\ 0, & \text{caso contrário}\end{cases} \quad (43)$$

- $\mathbb{H}(q,r)$ é o número mínimo de transições entre os estados q e r (Pode ser obtido pelo algoritmo de Floyd-Warshall, considerando que o peso de todas arestas é unitário), é importante lembrar que esse resultado será diferente do obtido pela distância de Hamming para os estados de fila. Por exemplo, a distância entre os estados $\{011a\}$ e $\{011b\}$ é igual a 2, e não 1;
- $\mathbb{Q}(q,r,l)$ é definida pela Equação (44):

$$\mathbb{Q}(q,r,l)=\begin{cases}1, & \text{caso } \mathbb{H}(q,r)=1, \mathbb{S}(q,k)=\mathbb{S}(r,k)\ \forall k, W(r,q,l)=1\\ 0, & \text{caso contrário}\end{cases} \quad (44)$$

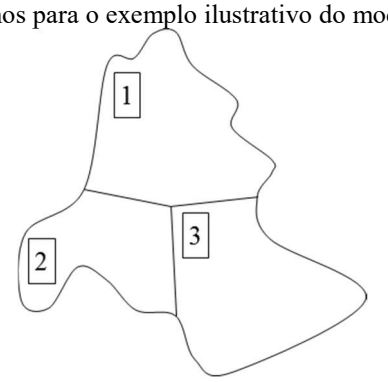
- $W(q,r,l)$ esta função indica o número de chamados da classe l a mais, ou a menos do estado q para o estado r; e
- $\mathbb{A}(q,k)$ conjunto de subátomos em que os servidores do grupo k são preferenciais enquanto no estado q (baseado na matriz de preferência de despacho).

4.2 Modelo hipercubo não-estacionário (com agregação de servidores)

Esta extensão tem por finalidade possibilitar analisar o desempenho de sistemas espacialmente distribuídos ao longo do tempo. Para tanto, desenvolve-se um modelo hipercubo que não está sujeito à hipótese de estacionariedade, a resolução desse tipo de problema é feita utilizando métodos numéricos apresentados no Anexo B. Para este trabalho em específico, utilizou-se o método de Runge-Kutta, por meio do software MATLAB® por meio da função 'ode45'. Para diminuição dos tempos computacionais, também já se considerou o modelo utilizando a agregação de servidores.

Para apresentar este modelo, é proposto um exemplo ilustrativo. A Figura 27 mostra um sistema composto por três átomos que opera ao longo e 12 horas. Os átomos não são alterados nesse período, permanecem com mesmo tamanho e numeração.

Figura 27 – Disposição dos átomos para o exemplo ilustrativo do modelo hipercubo não-estacionário.



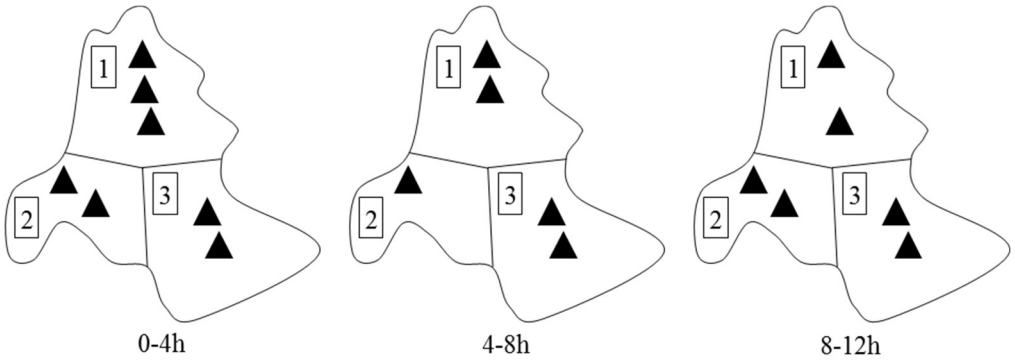
O processo de chegada dos chamados obedece um processo de Poisson não-homogêneo (Anexo D). Tal processo é resultado das probabilidades variáveis no tempo de um usuário entrar no sistema e é totalmente caracterizado por sua taxa de chegada. Uma forma comum é de modelar este processo é separar o tempo em intervalos com taxas constantes (BROWN *et al.*, 2005), a Tabela 5 mostra as taxas de chegada para os átomos ao longo das 12 horas de operação.

Tabela 5 – Taxas de chegada para o exemplo ilustrativo do modelo hipercubo não-estacionário.

Horário	Átomos		
	1	2	3
0-4h	0,8	0,4	0,5
4-8h	0,6	0,2	0,2
8-12h	0,5	0,3	0,4

Os servidores estão distribuídos entre os átomos e há mudanças de turno de quatro em quatro horas. A Figura 28 ilustra o número e a posição dos servidores ao longo do tempo. É importante ressaltar que a diferença no número de servidores de um turno para outro não representa necessariamente o número de servidores que param ou começam a trabalhar.

Figura 28 – Disposição dos servidores ao longo do tempo para o exemplo ilustrativo do modelo hipercubo não-estacionário.



Os servidores co-localizados são homogêneos e seus tempos de serviço não variam ao longo do tempo, pois considera-se que as condições de trânsito, treinamento das equipes e quaisquer variações operacionais são desprezíveis ao longo das 12 horas de operação do sistema. A Tabela 6 mostra as taxas de serviço para os servidores de cada átomo de acordo com suas localizações.

Tabela 6 – Taxas de serviço de acordo com a localização para o exemplo ilustrativo do modelo hipercubo não-estacionário.

Horário	Localização do servidor		
	Átomo 1	Átomo 2	Átomo 3
0-4h	1,5	1,6	1,4
4-8h	1,5	1,6	1,4
8-12h	1,5	1,6	1,4

O despacho dos servidores é simples e segue uma matriz de preferência de despacho fixa ao longo do tempo. Os servidores co-localizados também possuem as mesmas preferências de despacho entre si, tornando-os completamente indistinguíveis. Utilizando a agregação de servidores, é possível formar três grupos de servidores, um para cada átomo, respectivamente. Dessa maneira, no instante inicial, o sistema conta com 3 grupos, sendo o primeiro composto por 3 servidores, o segundo e o terceiro compostos por 2 servidores. Resumindo, neste instante

o sistema é representado por (3,2,2). A Tabela 7 traz a matriz de preferência de despacho utilizada pelos grupos de servidores para o atendimento.

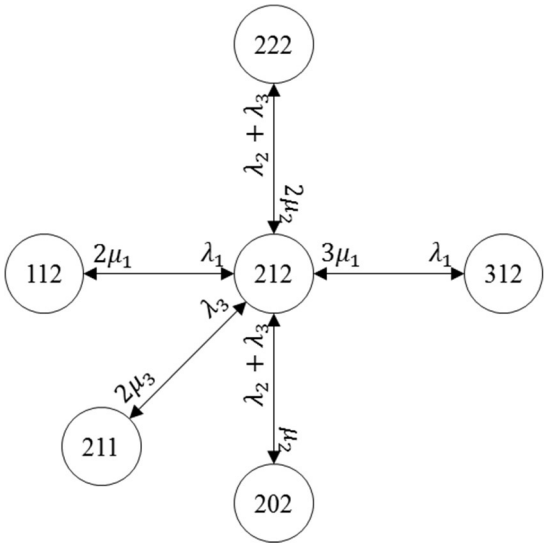
Tabela 7 – Matriz de preferência de despacho para o exemplo ilustrativo do modelo hipercubo não-estacionário.

Átomo	Grupo de servidores		
	1	2	3
1	1º	2º	2º
2	3º	1º	2º
3	3º	2º	1º

Para fins de modelagem, o sistema limita em 4 o número de máximo de usuários aceitos em fila. O espaço de estados do sistema pode ser representado da mesma maneira que o modelo estacionário. O Apêndice A traz o espaço de estados para o primeiro turno do sistema com suas transições e equações diferenciais.

Por se tratar de um modelo não-estacionário, as equações que descrevem o comportamento do sistema são equações diferenciais. O Anexo A mostra o processo de obtenção das equações diferenciais para um modelo *M*/*M*/*s*. Como o modelo hipercubo é uma expansão dos estados de um modelo *M*/*M*/*s*, a obtenção das equações diferenciais é direta. A Figura 29 ilustra o recorte no espaço de estados focado no estado {212} ao longo do primeiro turno (0-4h).

Figura 29 – Transições de estado no estado {212} para o exemplo ilustrativo do modelo hipercubo não-estacionário.



O fluxo de saída do estado pode ocorrer por 5 eventos. Primeiro, um dos servidores do grupo 1 finaliza um atendimento, levando ao estado {112}. Segundo, um dos servidores do

grupo 3 finaliza um atendimento, levando ao estado $\{211\}$. Terceiro, um servidor do grupo 2 finaliza um atendimento, levando ao estado $\{202\}$. Quarto, chega um chamado do átomo 1, levando ao estado $\{312\}$. Quinto, chega um chamado do átomo 2 ou do átomo 3, levando ao estado $\{222\}$.

O fluxo de entrada no estado também pode ocorrer por meio de 5 eventos. Primeiro, no estado $\{312\}$, um dos servidores do grupo 1 finaliza um atendimento. Segundo, no estado $\{222\}$, um dos servidores do grupo 2 finaliza um atendimento. Terceiro, no estado, $\{112\}$, chega-se um chamado do átomo 1. Quarto, no estado $\{211\}$, chega-se um chamado do átomo 3. Quinto, no estado $\{202\}$, chega-se um chamado do átomo 2 ou do átomo 3.

Entendendo os fluxos de entrada e saída do estado e, lembrando que o sistema ainda não atingiu o equilíbrio, há diferenças entre esses fluxos, pode-se determinar a Equação (45) diferencial que representa o comportamento neste estado.

$$\begin{aligned}\frac{dP_{\{212\}}(t)}{dt} &= -(\lambda(t) + 2\mu_1 + \mu_2 + 2\mu_3)P_{\{212\}}(t) + \lambda_1(t)P_{\{112\}}(t) + \lambda_3(t)P_{\{211\}}(t) \\ &\quad + \big(\lambda_2(t) + \lambda_3(t)\big)P_{\{202\}}(t) + 3\mu_1 P_{\{312\}}(t) + 2\mu_2 P_{\{222\}}(t)\end{aligned} \tag{45}$$

Como mencionado anteriormente, este problema trata-se de um Problema de Solução Inicial. Portanto, é preciso saber, ou estimar a situação inicial do sistema estudado. O exemplo ilustrativo começa a operar sem fila e com todos servidores desocupados, $P_{\{000\}}(0) = 1$.

Após quatro horas de operação, o sistema passa pela primeira troca de turno. Até este instante, o sistema é modelado utilizando uma rede de Markov de tempo contínuo. Para modelar este momento, utiliza-se uma rede de Markov de tempo discreto. Dessa maneira, pode-se classificar o modelo hipercubo não-estacionário como um modelo Misto de Redes de Markov de tempo discreto e contínuo (INGOLFSSON *et al.*, 2007). No entanto, para modelar a rede de Markov de tempo discreto, é preciso entender a disciplina de final de turno dos seus servidores.

O modelo apresentado até aqui é suficiente, caso o sistema não passe por uma troca de turno ou pausas para refeições, com mudanças no número de servidores. Caso contrário, deve-se considerar a disciplina de fim de turno adequada e, para isso, é necessário utilizar um modelo de cadeia de Markov de tempo discreto, como visto na Seção 3.3.2.

4.2.1 Modelo hipercubo considerando a disciplina exaustiva de fim de turno

Considerando que o exemplo ilustrativo é um sistema médico emergencial como o SAMU, os servidores são agora chamados de ambulâncias. A disciplina de final de turno

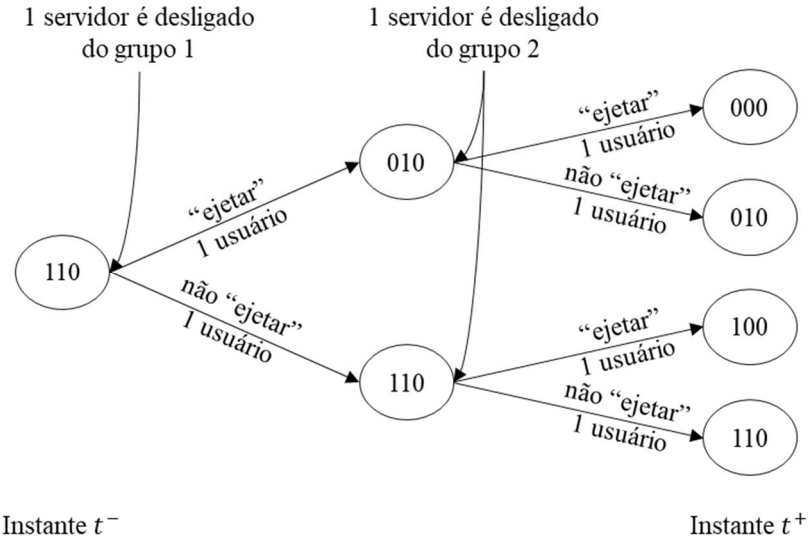
seguida pelo modelo é a disciplina exaustiva, as ambulâncias finalizam seus atendimentos antes de saírem do sistema.

Antes da mudança de turno, o sistema possui (3,2,2) servidores, ou seja, o vetor $s(t) = (3,2,2)$ para $t \in [0,4)$. No instante 4h, são desligados um servidor do grupo 1 e outro do grupo 2. Resumindo, são desligados (1,1,0) servidores, seguindo a mesma notação para servidores agrupados.

Assim como no modelo $M(t)/M/s(t)$ de Ingolfsson *et al.* (2007), visto na Seção 3.3, os usuários em atendimento por estes servidores desligados devem ser "ejetados" do sistema, visto que os servidores desligados no final do turno terminam o atendimento corrente e não realizam outros atendimentos *a posteriori*.

Considerando ainda o exemplo ilustrativo da Seção 4.2 e focando o estado $\{110\}$, sabendo que o número de servidores desligados na mudança de turno são dois servidores, um de cada grupo, 1 e 2 (1,1,0), pode-se encontrar o conjunto de possibilidades de transições instantâneas desse estado. A Figura 30 mostra as transições possíveis a partir do desligamento de (1,1,0) servidores, sendo que a taxa de transição é dada pelas probabilidades de se "ejetar" ou não um usuário devido ao desligamento do servidor. Por exemplo, a transição entre $\{010\}$ e $\{000\}$, após desligar um servidor do grupo 2, possui como taxa a probabilidade de se "ejetar" um usuário do grupo 2. Essas taxas são calculadas nos passos a seguir.

Figura 30 – Eventos possíveis para mudança de turno exaustiva no estado $\{110\}$ para o exemplo ilustrativo do modelo hipercubo não-estacionário.



A probabilidade de se "ejetar" um usuário se comporta como uma distribuição hipergeométrica. Assim como em Ingolfsson (2005), é possível fazer a analogia com uma urna,

considerando o estado $\{110\}$, onde o sistema possui (3,2,2) servidores e são desligados (1,1,0) servidores. Para a probabilidade de "ejetar" um usuário do grupo 1, o número total de bolas dentro da urna é o número de servidores do grupo 1 no instante antes da troca, $s_1(t^-) = 3$. O total de bolas brancas (sucessos) é o número de servidores do grupo 1 finalizando o turno, $\delta s_1 = 1$. O tamanho da amostra, retirada sem reposição, é o número de servidores ocupados no grupo 1, $n_1 = 1$. Por fim, a probabilidade de se retirar 1 bola branca da amostra, $\delta n_1 = 1$, é o mesmo que encontrar a probabilidade de se "ejetar" 1 usuário do grupo 1, como mostra a Equação (46). A Tabela 8 relaciona as notações dessa equação com os elementos do exemplo ilustrativo.

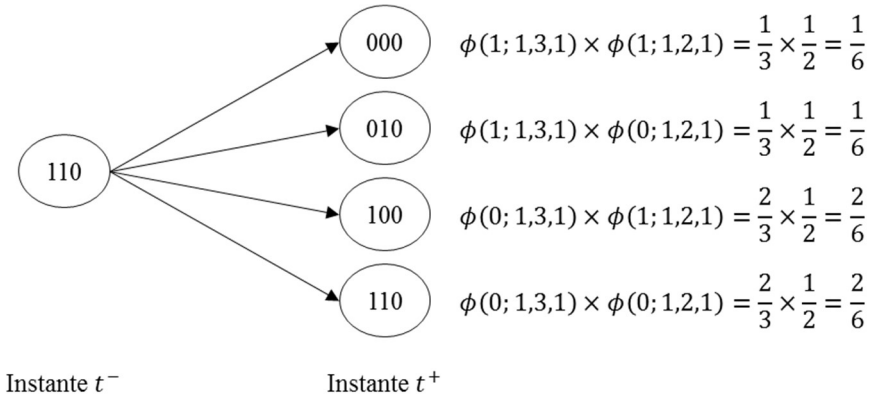
$$\phi(\delta n_1; \delta s_1, s_1(t^-), n_1) = \frac{\binom{n_1}{\delta n_1}\binom{s_1(t^-) - n_1}{\delta s_1 - \delta n_1}}{\binom{s_1(t^-)}{\delta s_1}} = \frac{\binom{1}{1}\binom{3-1}{1-1}}{\binom{3}{1}} = \frac{1}{3} \qquad (46)$$

Tabela 8 – Relação das notações utilizadas para a distribuição hipergeométrica durante a transição instantânea do modelo hipercubo não-estacionário.

Elemento representado	Exemplo ilustrativo	Notação
Estado do sistema	$\{110\}$	$\{n_1\ n_2\ n_3\}$
Servidores	(3,2,2)	$(s_1(t^-), s_2(t^-), s_3(t^-))$
Servidores desligados	(1,1,0)	$(\delta s_1, \delta s_2, \delta s_3)$
Usuários "ejetados"	-	δn_1, δn_2 e δn_3

Como a transição apresentada na Figura 30 é resultado de uma combinação de eventos, por exemplo, para que o sistema vá do estado $\{110\}$ para o estado $\{000\}$, é preciso "ejetar" um usuário do grupo 1 "e" "ejetar" um usuário do grupo 2. Isso quer dizer que essa probabilidade é resultado de o produto da probabilidade dos dois eventos ocorrerem, como ilustra a Figura 31. Por exemplo, a transição para o estado $\{010\}$ é resultado de $\phi(\delta n_1; \delta s_1, s_1(t^-), n_1) \times \phi(\delta n_2; \delta s_2, s_2(t^-), n_2)$.

Figura 31 – Probabilidades de transição instantânea do {110} para o exemplo ilustrativo do modelo hipercubo não-estacionário.



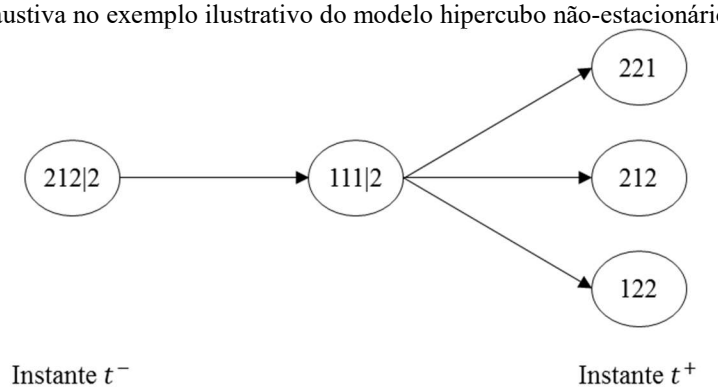
Este processo é utilizado para o cálculo do novo vetor de probabilidade usado como solução inicial para o período do exemplo ilustrativo que se inicia após 4 horas de operação. Contudo, após 8 horas de operação há uma nova mudança de turno. Nessa mudança de turno, novas ambulâncias entram em operação, uma no grupo 1 e outra no grupo 2. Por outro lado, uma das ambulâncias do grupo 1 é desligada, conforme mostrado na Figura 28.

No caso de novos servidores entrarem em atividade é preciso atentar-se para os chamados em fila. Com novos servidores começando a operar, caso haja chamados em fila, estes precisam ser distribuídos entre os servidores disponíveis. Como mencionado anteriormente, novos servidores podem entrar em atividade ao mesmo tempo em que outros são desligados e, por isso, é preciso também levar em conta os usuários "ejetados", conforme explicado nas Figura 30 e Figura 31.

Por exemplo, o estado $\{212|2\}$ representa a situação com dois usuários na fila. No instante que precede 8 horas de operação e a mudança de turno, o sistema possui (2,1,2) servidores. Como o sistema está completamente ocupado, a única transição possível, caso (1,0,1) servidores sejam desligados, é o sistema entrar em um estado $\{111|2\}$, por ter "ejetado" 1 usuário do grupo 1 e do grupo 3. Todavia, no instante que pospõe 8 horas de operação e a mudança de turno, o sistema possui (2,2,2) servidores, totalizando (1,1,1) servidores ainda livres para atender e ainda com 2 chamados em fila.

A Figura 32 ilustra o processo de "ejetar" os usuários no estado $\{212|2\}$ e depois distribuir os chamados em fila para os novos servidores disponíveis. Dessa maneira, as distribuições possíveis são: enviar um usuário para os grupos 1 e 2, ou enviar um usuário para os grupos 1 e 3, ou enviar um usuário para os grupos 2 e 3.

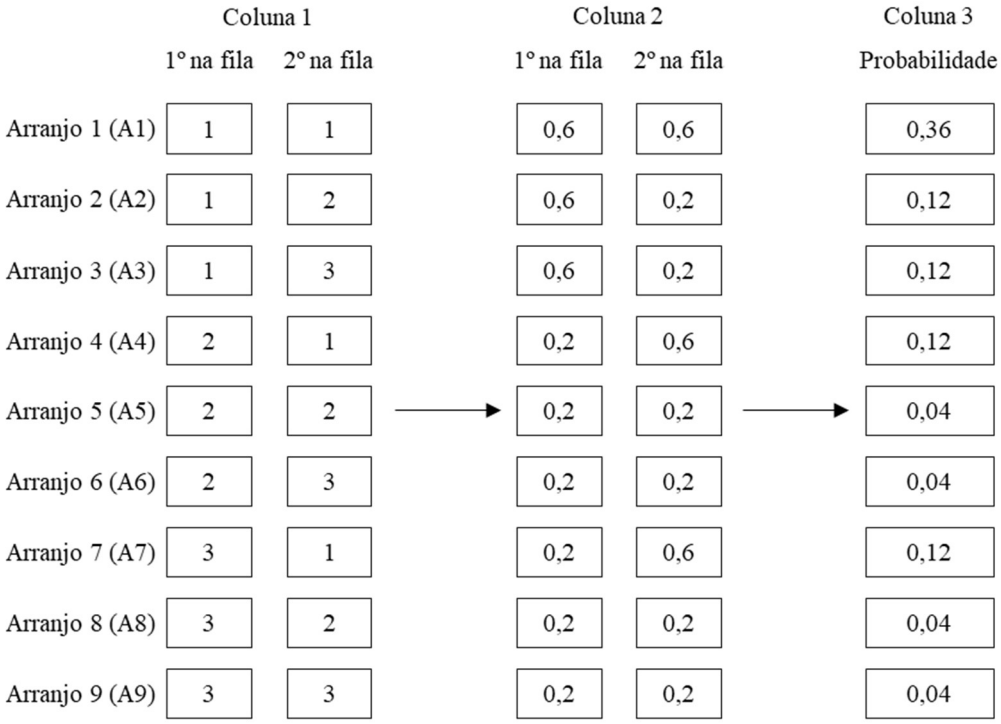
Figura 32 – Possibilidades de distribuição dos chamados da fila no estado $\{212|2\}$ para a mudança de turno exaustiva no exemplo ilustrativo do modelo hipercubo não-estacionário.



A probabilidade de cada grupo receber os usuários em fila depende da origem dos chamados, não diferenciados na fila como em Souza *et al.* (2015) ou Rodrigues *et al.* (2017), o que tornaria necessário saber se os servidores seriam capazes de receber os usuários. Sendo assim, uma fila simples com disciplina FCFS. Para conhecer probabilisticamente a origem dos chamados, é necessário o conhecimento das taxas de chegada no instante que precede a mudança do turno. Além da origem, é preciso conhecer a matriz (ou lista) de preferência de despacho obedecida pelo sistema logo após a mudança de turno. Em resumo, é preciso conhecer $\lambda_i(t^-)$ e a matriz de preferência de despacho em t^+.

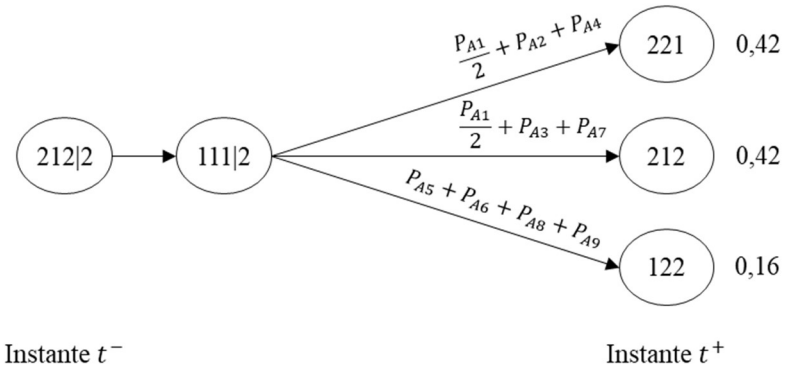
A Figura 33 mostra todos arranjos, já que a ordem dos chamados na fila é importante para distribuí-los aos novos servidores, para esses chamados em fila e suas probabilidades de ocorrência. A primeira coluna mostra as possíveis origens e ordem dos chamados em fila, sendo que os números de 1 a 3 representam seus átomos de origem. A segunda coluna traz a probabilidade de um chamado ser de seu respectivo átomo de origem, calculada por $\lambda_j(t^-)/\lambda(t^-)$. Por fim, na terceira coluna calcula-se a probabilidade de cada arranjo ocorrer multiplicando-se a probabilidade da origem dos chamados. É importante ressaltar que por este processo não se perde nenhum dado sobre a origem dos chamados e não utiliza nenhuma aproximação, portanto, a soma das probabilidades da terceira coluna é igual a 1.

Figura 33 – Arranjos possíveis para os chamados em fila no exemplo ilustrativo do modelo hipercubo não-estacionário.



Utiliza-se a matriz de preferência de despacho para encontrar qual dos estados possíveis da Figura 32 receberá cada um dos arranjos. A Figura 34 mostra a distribuição dos arranjos e o quanto cada estado receberá. Por exemplo, o Arranjo 3, seguindo a matriz de preferência de despacho (Tabela 7), o primeiro chamado da fila é enviado ao grupo 1 enquanto o segundo para o grupo 3, portanto, o terceiro arranjo leva ao estado $\{212\}$.

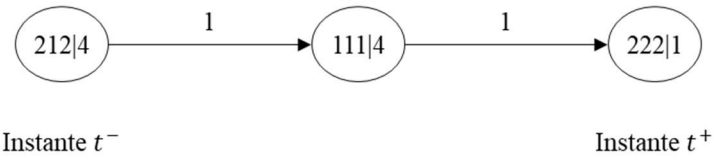
Figura 34 – Distribuição dos arranjos de chamados em fila para o exemplo ilustrativo do modelo hipercubo não-estacionário.



Ainda considerando a mudança de turno ocorrida às 8 horas, é possível que haja situações em que, mesmo após "ejetar" todos os usuários e redistribuir dos chamados em fila,

o sistema permanece saturado em t^+, como ilustra a Figura 35. Observe que primeiro ocorre a transição para o estado $\{111|4\}$, já que (1,0,1) servidores foram desligados e a mesma quantidade de usuários foi "ejetada", já que o sistema estava saturado. Após isso, em t^+ o sistema possui (2,2,2) servidores, um total de (1,1,1) servidores disponíveis, mas, como há 4 chamados em espera, independentemente da forma que os chamados em fila sejam distribuídos, o sistema permanece saturado e com 1 chamados ainda em fila. Por isso, a probabilidade do estados $\{212|4\}$ em t^- é integralmente transferida para o estado $\{222|1\}$ em t^+.

Figura 35 – Distribuição da probabilidade do estado $\{212|4\}$ para a mudança de turno exaustiva no exemplo ilustrativo do modelo hipercubo não-estacionário.



Os processos apresentados nesta seção mostram os valores não-nulos que compõe a matriz $B(t)$, matriz responsável pela transição instantânea nos momentos de mudança de turno apresentada para os modelos $M(t)/M/s(t)$, na Seção 3.3.2, ocorrendo conforme a Equação (25). O cálculo para os valores dessa matriz precisa ser generalizado o caso das Figuras 28 e 29, para o caso das Figuras 30 até 32 e para o caso da Figura 33.

A Equação (47) generaliza as situações em que não há chamados em fila (Figuras 29 e 30), apenas usuários são "ejetados". Dessa maneira, a transição apenas ocorre entre estados $\{n_1, n_2, n_3, 0\}$ e $\{n_1 - \delta n_1, n_2 - \delta n_2, n_3 - \delta n_3, 0\}$ (aqui optou-se por utilizar vírgulas na representação dos estados do sistema para facilitar a diferenciação dos grupos de servidores), pois ambos possuem 0 usuários em fila. Conforme mencionado anteriormente, na Seção 3.3.2, deve-se obedecer ao número máximo e mínimo de usuários "ejetáveis" por estado, como mostra a condição $\left(n_i - (s_i(t^-) - \delta s_i)\right)^+ \leq \delta n_i \leq \min(\delta s_i, n_i)$.

$$b_{\{n_1,n_2,n_3,0\},\{n_1-\delta n_1,n_2-\delta\ _2,n_3-\delta n_3,0\}} = \prod_{i=1}^{3} \phi(\delta n_i; \delta s_i, s_i(t^-), n_i)\,,$$
$$\text{para } n_i = 0, 1, \ldots, s_i(t^-) \quad \text{e } \left(n_i - (s_i(t^-) - \delta s_i)\right)^+ \leq \delta n_i \leq \min(\delta s_i, n_i)\,, \quad \forall i \tag{47}$$

A Equação (48) representa as situações em que o estado no instante t^- possui usuários na fila, mas após a mudança de turno haverá ao menos um servidor livre (Figuras 31 até 33). Dentre as condições é importante ressaltar a necessidade de haver fila de espera ($0 < Q$) e a

garantia de que todos usuários na fila serão distribuídos para os grupos de servidores ($Q = \sum_{i=1}^{3} \delta Q_i$). Além dessas condições ainda é preciso respeitar a quantidade mínima e máxima de usuários a serem "ejetados" ($(n_i - \delta n_i + \delta Q_i) \leq s_i(t^+)$ e $\left(n_i - (s_i(t^-) - \delta s_i)\right)^+ \leq \delta n_i \leq \min(\delta s_i, n_i)$).

$$b_{\{n_1,n_2,n_3,Q\},\{n_1-\delta n_1+\delta Q_1,n_2-\delta n_2+\delta Q_2,n_3-\delta n_3+\delta Q_1,0\}} = \left(\prod_{i=1}^{3} \phi(\delta n_i; \delta s_i, s_i(t^-), n_i)\right) P(\delta Q), \quad \text{para } 0 < Q = \sum_{i=1}^{3} \delta Q_i \ , (n_i - \delta n_i + \delta Q_i) \leq s_i(t^+) \text{ e } n_i = 0, 1, \dots, s_i(t^-) \text{ e } \left(n_i - (s_i(t^-) - \delta s_i)\right)^+ \leq \delta n_i \leq \min(\delta s_i, n_i)\,, \qquad \forall i \tag{48}$$

Em que:

- Q indica o número de usuários em fila antes da transição instantânea;
- δQ_i indica o número de usuários foram distribuídos da fila para o grupo i; e
- $P(\delta Q)$ representa a probabilidade de os usuários em fila serem distribuídos conforme o vetor δQ;

Por fim, a Equação (49) representa os casos representados pela Figura 35, em que, permanece em um estado saturado no instante seguinte. Para tanto é preciso obedecer a condição $\sum_{i=1}^{3}(n_i - \delta s_i) = -\delta Q + \sum_{i=1}^{3} s_i(t^+)$, onde após "ejetar" todos usuários necessários o sistema ainda terá mais usuários do que servidores operando, mesmo que novos servidores tenham entrado em operação.

$$b_{\{n_1,n_2,n_3,Q\},\{s_1(t^+),s_2(t^+),s_3(t^+),Q-\delta Q\}} = 1, \quad \text{para } \sum_{i=1}^{3}(n_i - \delta s_i) = -\delta Q + \sum_{i=1}^{3} s_i(t^+) \ e\,, \qquad \forall i \tag{49}$$

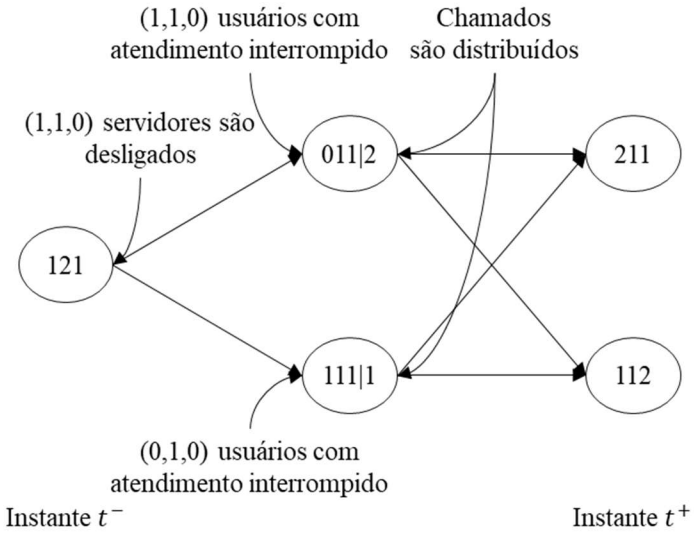
4.2.2 Modelo hipercubo considerando a disciplina preemptiva de fim de turno

Considere que o exemplo ilustrativo é uma operação de reparos para vias públicas após um período chuvoso, ou após o degelo da primavera. Ao final de um turno, as equipes interrompem o serviço, deixando a rua sinalizada e esta volta para a fila de serviços. Usuários não são "ejetados" do sistema, eles voltam para a fila e são redistribuídos para os servidores disponíveis.

Ainda considerando o exemplo ilustrativo da Seção 4.2, o sistema conta com (3,2,2) servidores no início da operação. Após 4 horas de operação, (1,1,0) servidores são desligados interrompendo seus serviços. Esses serviços interrompidos precisam ser redistribuídos entre os servidores disponíveis ou ficar na fila.

Focando no estado $\{121\}$, sabendo que (1,1,0) servidores serão desligados e que o próximo período opera com (2,1,2) servidores, é possível encontrar as possibilidades de transições instantâneas. A Figura 36 mostra as possíveis transições instantâneas encontradas em dois passos. O primeiro passo consiste em encontrar o conjunto de estados intermediários possíveis por meio das interrupções nos serviços. Os serviços interrompidos entram em uma fila transitória, como os estados intermediários $\{011|2\}$ e $\{111|1\}$ com 2 e 1 usuários nesta fila, respectivamente. O segundo passo é encontrar o conjunto de estados finais possíveis de serem atingidos por meio da distribuição desses chamados em fila. Por exemplo, no estado $\{121\}$, já que o sistema opera com (3,2,2) servidores e (1,1,0) servidores serão desligados, o grupo 2 está com todos seus servidores ocupados e, obrigatoriamente, 1 de seus usuários terá seu atendimento interrompido. Por outro lado, apenas 1 servidor está ocupado no grupo 1, caso este servidor não seja desligado e seu serviço interrompido, o sistema entra no estado intermediário $\{111|1\}$. O chamado na fila transitória do estado $\{111|1\}$ pode ser distribuído, conforme seu átomo e origem e a matriz de preferência de despacho (Tabela 7), para um servidor disponível do grupo 1, ou do grupo 3, levando aos estados $\{211\}$ ou $\{112\}$, respectivamente. Neste exemplo o grupo 2 não pode receber os chamados da fila transitória nos estados $\{111|1\}$ e $\{011|2\}$, porque já se encontra saturado.

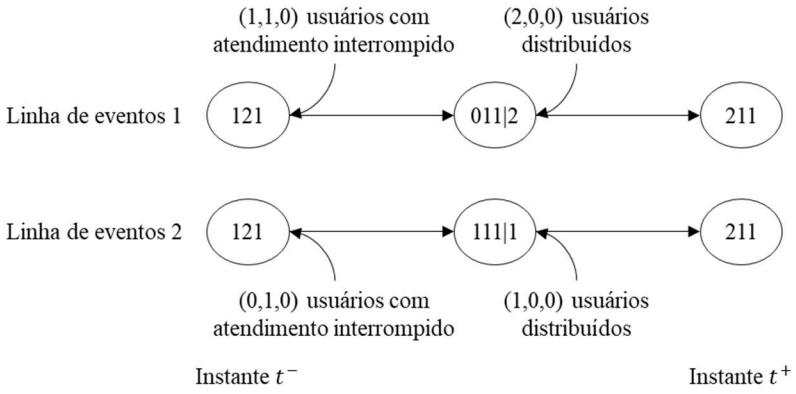
Figura 36 – Possibilidades de distribuição dos chamados interrompidos no estado {121} para a mudança de turno preemptiva no modelo hipercubo não-estacionário.



O cálculo da probabilidade de interrupção de serviços em cada grupo é idêntico à disciplina exaustiva, segue uma distribuição hipergeométrica. A distribuição corresponde à busca por δn_i bolas brancas (sucessos) em uma urna com uma população total de $s_i(t^-)$ bolas, sendo δs_i delas brancas, o tamanho da amostra é n_i retirada sem reposição. A Equação (46) da Seção 4.2.1 representa esta situação.

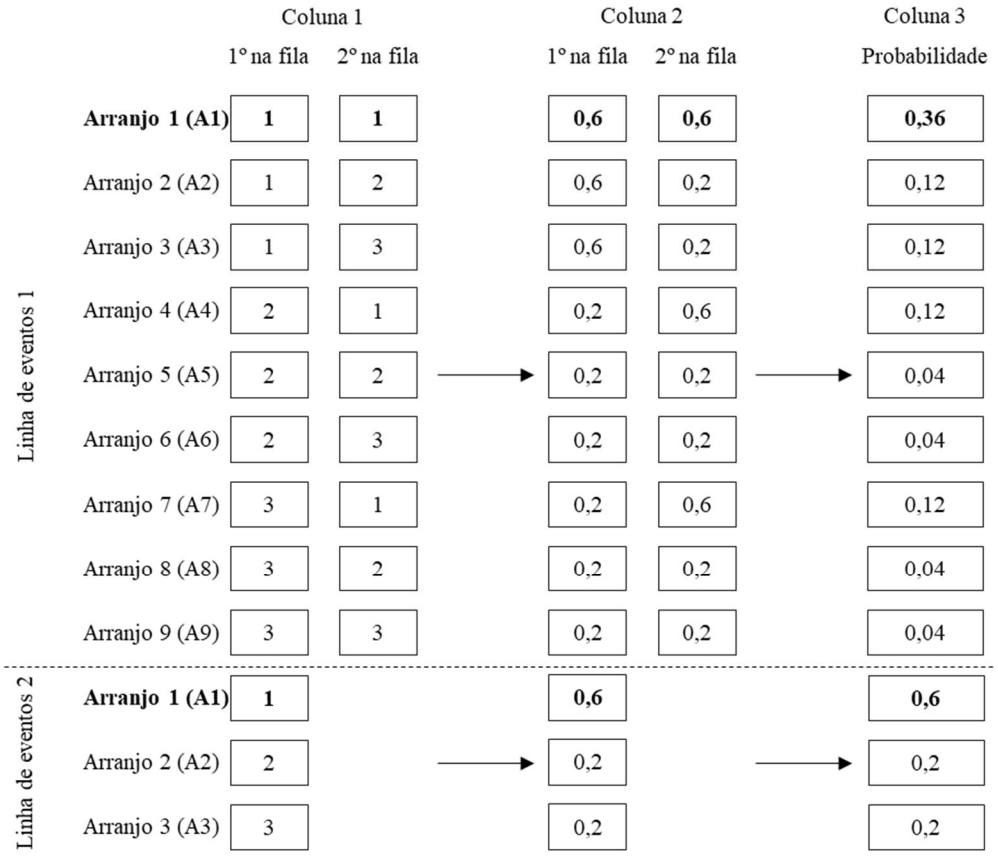
A distribuição dos chamados é feita encontrando o conjunto de arranjos que os chamados interrompidos podem formar, assim como os chamados em fila na disciplina exaustiva. Além disso, como mostra a Figura 37, pode existir mais de uma combinação de eventos que leva do estado $\{121\}$ para o estado $\{211\}$. O exemplo mostra dois recortes da transição vista na Figura 36 em que é possível interromper (1,1,0) serviços e ambos serem atendidos pelo grupo 1 (primeira linha de eventos), ou interromper apenas (0,1,0) serviço e este também ser atendido pelo grupo 1 (segunda linha de eventos), em ambos casos sai-se do estado $\{121\}$ e chega no estado $\{211\}$.

Figura 37 – Combinações de eventos possíveis entre os estados $\{121\}$ e $\{211\}$ para a disciplina preemptiva no modelo hipercubo não-estacionário.



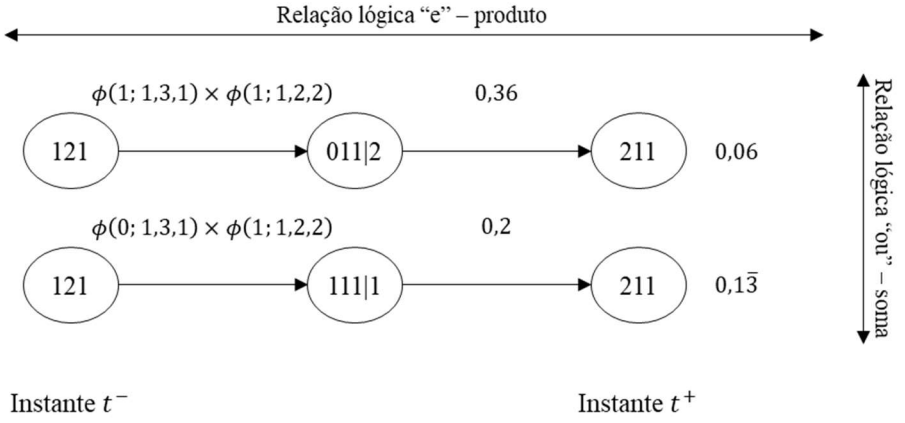
A distribuição dos chamados interrompidos é feita considerando todos os arranjos possíveis que os chamados podem formar durante a fila transitória, assim como as taxas de chegada em t^-, a matriz de preferência de despacho em t^+ (Tabela 7) e o número de servidores operando em t^+, semelhantemente à disciplina exaustiva. A Figura 38 ilustra as probabilidades dos arranjos dos chamados na fila transitória para as duas linhas de eventos. Os arranjos que levam ao estado $\{211\}$, em cada linha de eventos, foram destacados em negrito.

Figura 38 – Arranjos e suas probabilidades para os casos de 1 e 2 chamados interrompidos na disciplina preemptiva no modelo hipercubo não-estacionário.



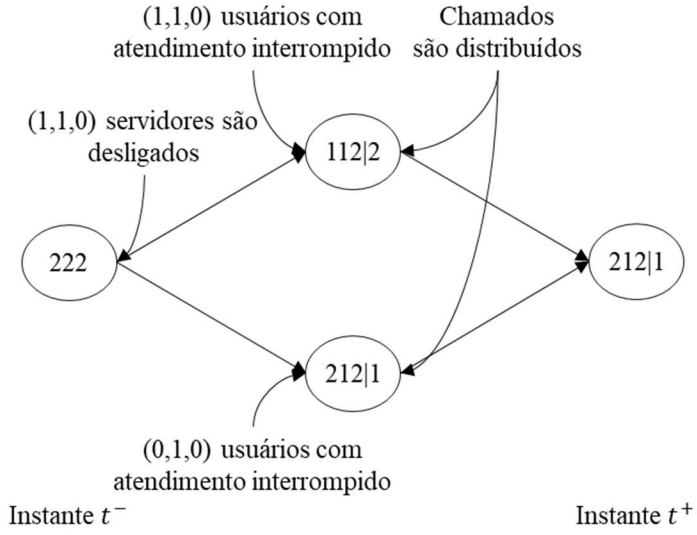
A Figura 39 mostra a probabilidade total de o sistema sofrer a transição instantânea entre os estados $\{121\}$ e $\{211\}$. Além disso, é ilustrado que os eventos que ocorrerem horizontalmente são sequenciais e, portanto, possuem uma relação lógica "e", enquanto verticalmente são os eventos que podem ocorrer em paralelo, possuindo uma relação lógica "ou". Com isso, a probabilidade total de o sistema sair do estado $\{121\}$ e entrar no estado $\{211\}$, seguindo uma disciplina preemptiva, somando as duas linhas de eventos, é de $0{,}19\overline{3}$.

Figura 39 – Probabilidades de transição entre os estados {121} e {211} na disciplina preemptiva para o modelo hipercubo não-estacionário.



Ainda considerando a mudança de turno do exemplo ilustrativo da Seção 4.2 após 4 horas de operação, focando no estado {222} encontra-se a transição instantânea representada na Figura 40. Observe que como o número de usuários no sistema antes da transição era maior do que a capacidade do sistema após a transição, invariavelmente o sistema ficou saturado, onde antes havia 1 servidor disponível, agora há 1 usuário em espera.

Figura 40 – Possibilidades de distribuição dos chamados interrompidos no estado {222} para a mudança de turno preemptiva no modelo hipercubo não-estacionário.



O processo de transição instantânea apresentado entre as Figuras 34 e 37 pode ser resumido pela Equação (50), situações em que haverá ao menos um servidor disponível para atendimento após a transição ($Q + \sum_{i=1}^{3} n_i < \sum_{i=1}^{3} s_i(t^+)$). Observe que uma das condições para aplicar esta equação é que o número de serviços interrompidos deve ser igual ao número de serviços redistribuídos ($\sum_{i=1}^{3} \delta p_i = \sum_{i=1}^{3} \delta n_i$). Também precisa-se respeitar o número

mínimo e máximo de usuários a terem seus serviços interrompidos ($\left(n_i - (s_i(t^-) - \delta s_i)\right)^+ \leq \delta n_i \leq \min(\delta s_i, n_i)$). Por fim, também é necessário garantir que todos os chamados em fila serão distribuídos entre os grupos de servidores ($Q = \sum_{i=1}^{3} \delta Q_i$)

$$b_{\{n_1,n_2,n_3,Q\},\{n_1-\delta n_1+\delta p_1+\delta Q_1,n_2-\delta n_2+\delta p_2+\delta Q_2,n_3-\delta n_3+\delta p_3+\delta Q_3,0\}} = \sum_{-\delta n+\delta p+\delta Q=} \left(\left(\prod_{i=1}^{3} \phi(\delta n_i; \delta s_i, s_i(t^-), n_i)\right) P(\delta p + \delta Q)\right),$$

$$\text{para } n_i = 0, 1, \dots, s_i(t^-) \quad \text{e } \left(n_i - (s_i(t^-) - \delta s_i)\right)^+ \leq \delta n_i \tag{50}$$

$$\leq \min(\delta s_i, n_i), Q + \sum_{i=1}^{3} n_i < \sum_{i=1}^{3} s_i(t^+), \sum_{i=1}^{3} \delta p_i = \sum_{i=1}^{3} \delta n_i, 0 \leq Q = \sum_{i=1}^{3} \delta Q_i, \quad \forall i$$

Em que:

- δp_i indica o número de serviços interrompidos que foram redistribuídos para o grupo i; e
- $P(\delta p + \delta Q)$ representa a probabilidade de os chamados interrompidos e os chamados em fila sejam distribuídos conforme os vetores δp e δQ, respectivamente;

Seguindo com o caso da Figura 40, em que o sistema ficará saturado após a transição instantânea, é resumido pela Equação (51). Uma condição é que o sistema possua mais usuários no sistema em t^- do que servidores operando em t^+ ($Q + \sum_{i=1}^{3} n_i - \sum_{i=1}^{3} s_i(t^+) = \delta Q \geq 0$). Neste caso é importante lembrar que, na disciplina preemptiva, o sistema mantém o número de usuários no sistema. Isso apenas não ocorre caso δQ seja maior do que a capacidade máxima do sistema, $\max Q$, resultando em perda para o sistema.

$$b_{\{n_1,n_2,n_3,Q\},\{s_1(t^+),s_2(t^+),s_3(t^+),K\}} = 1,$$

$$\text{para } n_i = 0, 1, \dots, s_i(t^-), Q + \sum_{i=1}^{3} n_i - \sum_{i=1}^{3} s_i(t^+) = \delta Q \geq 0, \tag{51}$$

$$K = \min(\delta Q, \max Q)$$

Em que:

- δQ representa o número de usuários que entrariam na fila após a interrupção dos serviços;
- $\max Q$ é a capacidade máxima de usuários em fila aceita pelo sistema; e
- K é o número de usuários que, de fato, entraram na fila (não foram perdas do sistema).

4.2.3 Hipóteses para o modelo hipercubo não-estacionário

A aplicação do modelo hipercubo não-estacionário requer que as hipóteses do modelo clássico sejam revistas. O exemplo ilustrativo contempla as hipóteses.

i) Existência de átomos geográficos: a região onde são prestados os serviços deve ser dividida em N_A átomos geográficos, sendo que cada átomo corresponde a uma fonte independente de chamados. Os átomos não podem sofrer alterações geográficas ao longo do tempo, como partições e aglutinações;
ii) Processo de chegada: deve ser um processo de Poisson heterogêneo. Os usuários de cada átomo solicitam chamados por meio do processo de Poisson, sendo os chamados independentes entre si. Além disso, as funções de taxa de chegada, $\lambda_j(t)$, de chamados de cada átomo deve ser conhecida ao longo de todo o período de análise;
iii) Tempos de viagem aos átomos: a função dos tempos de viagem $\tau_{ij}(t)$ de cada átomo i para o átomo j deve ser conhecida ou estimada para todo o período de análise;
iv) Servidores do sistema: existe um vetor $s(t)$ que representa o número de servidores espacialmente distribuídos ao longo do sistema, ao longo do período de análise. Todos os servidores são capazes de se deslocar e atender a qualquer um dos átomos. Também existe um vetor $\delta s(t)$ que representa o número de servidores finalizando seu turno no instante t. A troca de turno deve seguir uma disciplina bem definida de acordo com a operação do sistema, basicamente uma disciplina preemptiva ou exaustiva;
v) Localização dos grupos de servidores: a localização dos grupos de servidores deve ser conhecida ao menos probabilisticamente ao longo do período de análise. Os servidores podem se mover pelos átomos, localização probabilística, ou ficar fixo em um deles;
vi) Despachos dos servidores: para atender a qualquer chamado é enviado apenas um servidor para o local. Se não houverem servidores disponíveis, os chamados entram em fila, com disciplina *FCFS*, ou são considerados perdas do sistema;
vii) Política de despacho dos servidores: para todos os instantes de tempo há uma lista (ou matriz) de preferência de despacho para cada átomo, obedecendo uma ordem de envio dos servidores para os chamados que pode ser alterada ao longo do tempo;
viii) Tempo de serviço: o tempo de serviço de um servidor engloba o tempo de setup, o tempo de viagem e o tempo em cena até o retorno à base (ou área) de origem. Os tempos de serviço devem ser exponencialmente distribuídos e não variam ao longo do período de análise; e

ix) Dependência do tempo de serviço em relação ao tempo de viagem: a variação do tempo de viagem deve ser considerada uma variável de segunda ordem no tempo total de serviço, quando comparado ao tempo em cena e preparação da equipe.

Além das nove hipóteses adaptadas do modelo clássico, é preciso que uma décima hipótese seja acrescentada para que seja possível resolver o sistema de equações diferenciais do sistema.

x) Situação inicial do sistema: é preciso conhecer ou estimar probabilisticamente a situação inicial do sistema por meio de um vetor de probabilidade $\pi(0)$. Por exemplo, o sistema pode se encontrar vazio no início da operação, ou o vetor de probabilidade pode ser resultado de uma mudança de turno, ou pode estimar-se que o sistema comece a operação em um regime estacionário.

Declarar as hipóteses de aplicação do modelo é um passo importante para permitir que novas extensões sejam criadas. Junto a isso, pode-se utilizar as informações de forma a possibilitar o seu uso em rotinas de otimização. Por exemplo, alterando a matriz de preferência de despacho, reorganizar e redistribuir os turnos dos servidores, incluir calibração dos tempos de serviço, etc.

Neste sentido, é importante que se ressalte que as distribuições dos chamados apresentada nas disciplinas exaustiva e preemptiva são modelos em que o tomador de decisão é totalmente racional, seguindo rigorosamente a matriz de preferência de despacho e as informações sobre as taxas de chegada. Este caso não é obrigatório, a distribuição pode ser feita conforme o tomador de decisão bem entender, desde que obedeça aos conceitos de o porquê "ejetar" usuários na disciplina exaustiva e redistribuir os chamados na preemptiva, distribuindo os chamados integralmente.

4.2.4 Medidas de desempenho para o modelo hipercubo não-estacionário

Para concluir a análise do exemplo ilustrativo, o modelo possui medidas de desempenho interessantes para o gerenciamento do sistema, quando satisfeitas as hipóteses. As medidas de desempenho são tanto em nível de servidor quanto em nível dos átomos e subátomos, ou o sistema como um todo.

Para ilustrar as diferenças entre as disciplinas de fim de turno, os resultados para cada disciplina recebem uma denominação de acordo com o tipo de sistema que as disciplinas costumam representar melhor. A disciplina exaustiva é denominada por "Ambulâncias", enquanto a disciplina preemptiva é denominada por "Tapa buracos".

4.2.4.1 Nível de serviço

Utilizando as definições de nível de serviço para o modelo M(t)/M/s(t) encontradas em Green e Soares (2007) e Ingolfsson *et al.* (2007), pode-se adaptá-las para o modelo hipercubo não-estacionário. A Equação (52) mostra a adaptação. Em casos de aumento no número de servidores, considera-se os estados que após a transição instantânea, siga ela a disciplina exaustiva ou preemptiva, estarão saturados, situações das Equações (49) e (51), respectivamente. Sendo válida a expressão $s(t) + \delta s = s(t + \epsilon)$, com $\tau > \epsilon$. É preciso obedecer, visto que a solução do problema é numérica, o limite de capacidade da fila K.

$$SL(t) = \begin{cases} 1 - \sum_{i=0}^{K} p_{\{s(t),i\}} \sum_{j=0}^{i} \frac{a^{-j} e^{-a}}{j!}, & s(t+\tau) \leq s(t) \\ 1 - \sum_{i=0}^{K-\delta} p_{\{s(t)+\delta\ ,i+\delta Q\}} \sum_{j=0}^{i} \frac{a^{-j} e^{-a}}{j!}, & s(t+\epsilon) = s(t) + \delta s \end{cases} \tag{52}$$

Para calcular o número médio de saídas, $a(t, \tau)$, é necessário somar o número médio de saídas de cada grupo do modelo, como mostra a Equação (53). Sendo que $a_k(t, \tau)$ é o número médio de saídas do grupo k.

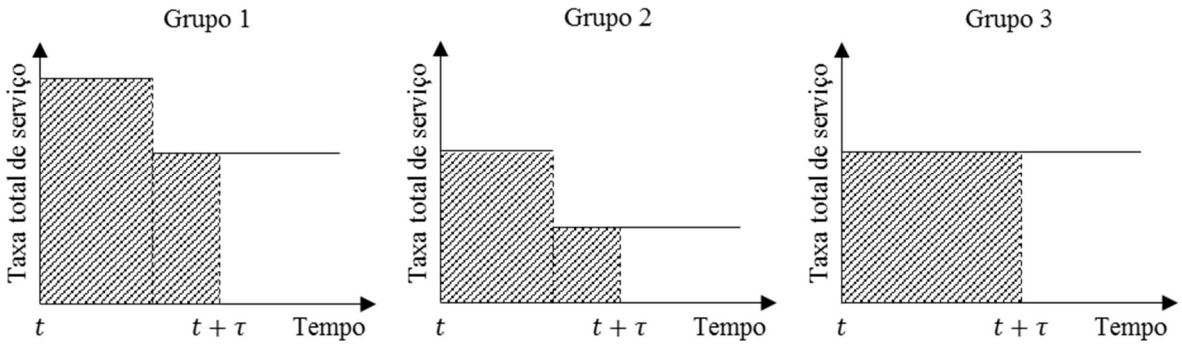
$$a(t, \tau) = \sum_{k=1}^{3} a_k(t, \tau) \tag{53}$$

O número médio de saídas de cada grupo pode ser calculado de forma análoga ao modelo *M*(*t*)/*M*/*s*(*t*). A Equação (54) mostra o cálculo para cada situação, considerando que há apenas uma troca de turno ao longo do período $(t, t + \tau]$.

$$\begin{aligned} a_k(t, \tau) &= \mu_k \tau s_k(t), && \text{para } s_k(t) = s_k(t + \tau) \\ a_k(t, \epsilon, \tau) &= \mu_k\big(\epsilon s_k(t) + (\tau - \epsilon)(s_k(t) + \delta s_k)\big), && \text{para } s_k(t) < s_k(t + \tau) \\ a_k(t, \epsilon, \tau) &= \mu_k\big(\epsilon s_k(t) + (\tau - \epsilon)(s_k(t) - \delta s_k)\big), && \text{para } s_k(t) > s_k(t + \tau) \end{aligned} \tag{54}$$

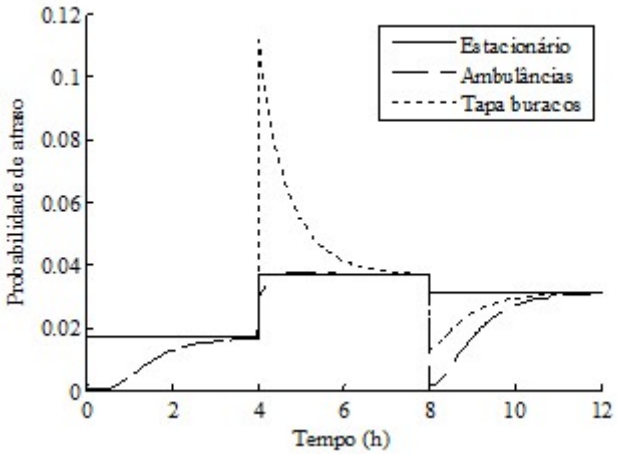
Esse cálculo pode ser observado instantes antes da troca de turno após 4 horas de operação no exemplo ilustrativo. A Figura 41 mostra as alterações na taxa total de serviço quando há a troca de turno, por meio da manutenção dos servidores no grupo 2 e a redução de um servidor nos grupos 1 e 2.

Figura 41 – Cálculo do número médio de saídas em cada grupo do exemplo ilustrativo do modelo hipercubo não-estacionário.



Por fim, a Figura 42 ilustra a probabilidade de atraso do exemplo ilustrativo, comparando os resultados do modelo estacionário, da disciplina exaustiva (Ambulâncias) e da disciplina preemptiva (Tapa buracos). Lembrando que a probabilidade de atraso é o complemento do nível de serviço com $\tau = 0$. Os desvios com relação ao modelo estacionário podem ser próximos a 250%, como visto no instante 4h. Na disciplina exaustiva, os desvios chegam à 100% em 0h e 8h, ficando sempre a baixo do valor para o estacionário.

Figura 42 – Comparação da probabilidade de atraso no exemplo ilustrativo do modelo hipercubo não-estacionário.



4.2.4.2 Tempo médio de espera

A definição de tempo médio de espera apresentada por Green e Soares (2007), na Equação (34), é expandida para calcular o tempo médio de espera de forma aproximada. Para tanto, será definido o cálculo para o termo $E\left(W_Q^i(t)\right)$ de forma aproximada.

Lembrando que os tempos de serviço são exponencialmente distribuídos o tempo médio para a próxima pessoa ser atendida é dado pelo valor esperado de uma variável aleatória

exponencialmente distribuída (GREEN; SOARES, 2007). A variável aleatória corresponde ao número de serviços que o novo usuário precisa esperar serem completos para que seja atendido, como mostra a Equação (55). Sendo que $n_{Q\{s(t),k\}}$ representa o número de usuários em fila no estado $\{s(t), k\}$, com todos servidores ocupados e k usuários em fila. A taxa média de serviço percebida pelo usuário é representada por $\bar{\mu}(t)$.

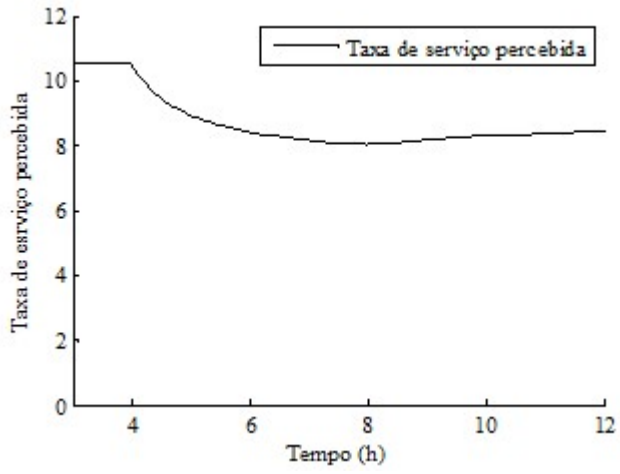
$$E\left(W_Q^{\{s(t),k\}}\right) = \frac{n_{Q\{s(t),k\}} + 1}{\bar{\mu}(t)}, \qquad \text{para } k = 0, 1, 2, \dots, K \tag{55}$$

Para se calcular uma aproximação da taxa de serviço média, $\bar{\mu}(t)$, é preciso entender os eventos futuros ao instante t e suas probabilidades de afetarem a taxa de serviço percebida pelo usuário. A percepção do usuário do sistema a respeito da taxa média de serviço varia ao longo do tempo em que o usuário está no sistema. Isso porque o conforme mudam as taxas de serviço com a entrada ou saída de servidores o número médio de saídas do sistema altera-se. Assim, a taxa de serviço percebida por um usuário a partir do instante t é resultado do número médio de saídas durante o lapso de tempo até $t + i\tau$ e do próprio intervalo de tempo $i\tau$, como mostra a Equação (56). Sendo que i é o número de observações realizadas a cada intervalo de tempo τ.

$$Taxa\ de\ serviço\ percebida\ pelo\ usuário = \frac{a(t, i\tau)}{i\tau}, \qquad \text{para } \forall i \in \mathbb{Z}_+ \tag{56}$$

Por exemplo, a Figura 43 ilustra a percepção da taxa de serviço de um usuário que chega ao sistema do exemplo ilustrativo em um instante próximo a $t = 3$ horas de operação. É possível perceber que a taxa média de serviço percebida começa a diminuir após 4h, momento de troca de turno com a saída de dois servidores. A taxa média de serviço volta a crescer após 8h, quando um novo servidor começa a operar.

Figura 43 – Taxa de serviço percebida por um usuário a partir de 3h no exemplo ilustrativo do modelo hipercubo não-estacionário.

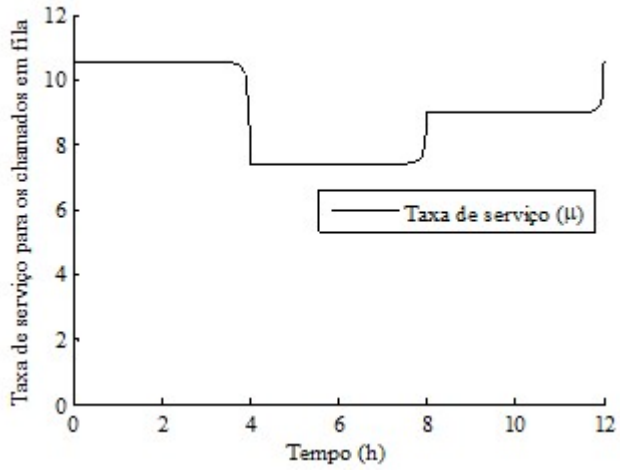


Contudo, a taxa de serviço com que o usuário é atendido deve ser ponderada, já que o usuário não deve ficar um tempo $i\tau$ indefinido no sistema. O tempo de serviço é exponencialmente distribuído, logo os eventos que ocorrem próximos ao instante t afetam mais o atendimento do que ocorrem horas mais tarde. Por isso, para cada intervalo de tempo τ é atribuído um peso exponencial, correspondente ao acréscimo da curva exponencial acumulada no intervalo $(t + (i-1)\tau, t + i\tau]$. Como resultado, obtém-se a aproximação da taxa de serviço média, $\bar{\mu}(t)$, na Equação (57). A taxa de serviço inicial utilizada na ponderação é $\mu_0 = s(t)\mu$. Para casos em que a taxa de serviço sofre variações muito grandes, é possível realizar um processo iterativo $\bar{\mu}^k(t) \rightarrow \mu_0^{k+1}$. Lembrando que, quanto menor for o intervalo τ, maior é a precisão de $\bar{\mu}(t)$.

$$\bar{\mu}(t) = \sum_{i=1}^{\infty} \frac{a(t, i\tau)}{i\tau}\left(e^{-\mu_0(i-1)\tau} - e^{-\mu_0 i\tau}\right), \qquad \text{para } \forall i \in \mathbb{Z}_+ \tag{57}$$

A Figura 44 mostra a aproximação da taxa de serviço para o exemplo ilustrativo. Próximo aos instantes 4h, 8h e 12h há uma mudança na taxa de serviço. Isto quer dizer que os usuários que chegam em horários próximos às mudanças de turnos têm seu tempo de espera afetado por elas. Assim, momentos antes da saída de um servidor a taxa de serviço começa a decrescer e o tempo de espera a aumentar.

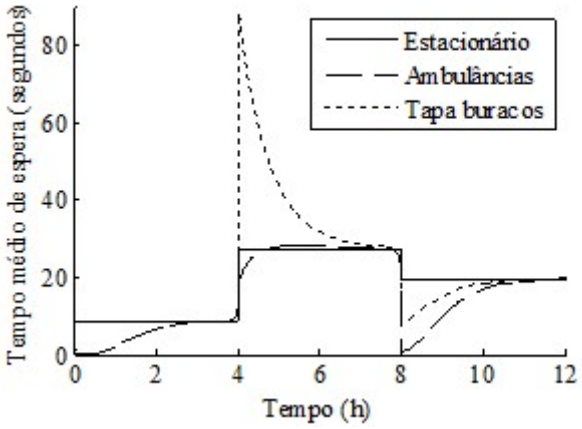
Figura 44 – Evolução da taxa de serviço para os chamados em fila no exemplo ilustrativo do modelo hipercubo não-estacionário.



Como mostra a Figura 45 a escolha das disciplinas de fim de turno tem bastante impacto sobre os tempos médios de espera do sistema. Assim como o nível de serviço, o desvio-relativo em comparação ao modelo estacionário pode chegar à 250% na disciplina preemptiva e 100%

na disciplina exaustiva. Além disso, é possível notar os efeitos de se desligar um servidor momentos antes do evento, como sugerido na Figura 44.

Figura 45 – Tempo médio de espera para o exemplo ilustrativo do modelo hipercubo não-estacionário.



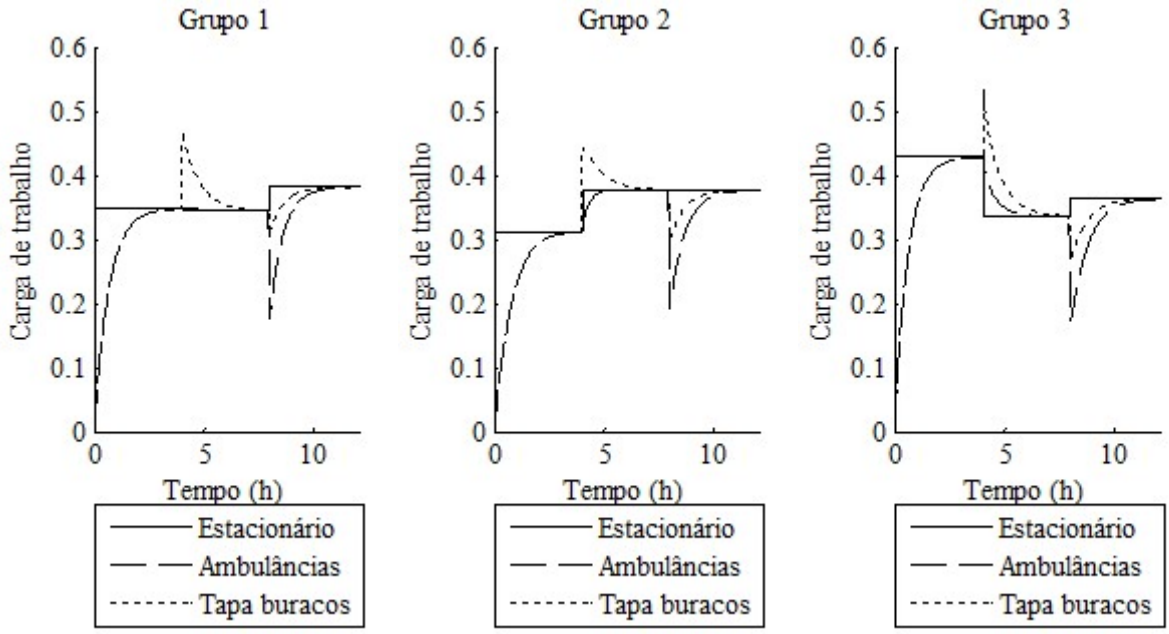
4.2.4.3 Carga de trabalho (Workload)

A carga de trabalho, como dependente diretamente das probabilidades de estado instantâneas, tem em seu cálculo uma forma direta proveniente do modelo estacionário. Como mostra a Equação (58). Note que o número de servidores ocupados no estado B independe do tempo.

$$\rho_{ki}(t) = \sum_{B \in M} \frac{n_{kB} \cdot P_B(t)}{n_k(t)}, \qquad \text{para } n_k > 0 \text{ e } \forall t \tag{58}$$

Para o exemplo ilustrativo, a Figura 46 ilustra a evolução das cargas de trabalho dos servidores de cada grupo. É possível observar o mesmo comportamento, quanto às disciplinas preemptiva e exaustiva, observado no nível de serviço, sendo que a disciplina exaustiva ficou em todos instantes com cargas inferiores ou iguais à disciplina preemptiva.

Figura 46 - Carga de trabalho dos servidores de cada grupo para o exemplo ilustrativo do modelo hipercubo não-estacionário.



4.2.4.4 Frequência de despacho

Assim como a carga de trabalho, o cálculo da frequência de despacho vem diretamente do seu cálculo no modelo estacionário, como mostra a Equação (59). Note que a probabilidade de envio de um servidor é variável no tempo devido à variação das probabilidades de estado e pelas possíveis mudanças na matriz de preferência de despacho.

$$f_{ki,j}(t) = f_{ki,j}^{(nq)}(t) + f_{ki,j}^{(q)}(t)$$

$$f_{ki,j}^{(nq)}(t) = \frac{\lambda_j(t)}{\lambda(t)} \cdot \frac{\mu_{ki}(t)}{\mu_k(t)} \cdot \sum_{D \in E_{k,j}(t)} P_D(t) \tag{59}$$

$$f_{ki,j}^{(q)} = \frac{\lambda_j(t)}{\lambda(t)} \cdot P_S \cdot \frac{\mu_{ki}(t)}{\mu(t)}$$

4.2.4.5 Tempos médios de viagem

Os tempos médios de viagem do sistema podem ser estimados a partir da função dos tempos médios de viagem entre os átomos, $\tau_{ij}(t)$. A Tabela 9 mostra os valores para o exemplo ilustrativo, que são invariáveis no tempo ($\tau_{ij}(t) = \tau_{ij}$).

Tabela 9 – Tempos médios de viagem (minutos) entre os átomos do exemplo ilustrativo para o modelo hipercubo não-estacionário.

Átomos	1	2	3

1	5	10	15
2	10	5	10
3	15	10	5

A partir das localizações dos grupos de servidores, pode-se encontrar os tempos médios de viagem de cada um desses para os átomos do sistema, t_{kj}. O cálculo pode ser realizado utilizando a mesma Equação (4) do modelo hipercubo clássico. A Tabela 10 mostra os valores para o exemplo ilustrativo, também invariáveis no tempo, visto que a localização dos grupos de servidores e os tempos médios de viagem não variam ao longo do tempo.

Tabela 10 – Tempos médios de viagem (minutos) de cada grupo de servidores para os átomos do sistema no exemplo ilustrativo do modelo hipercubo não-estacionário.

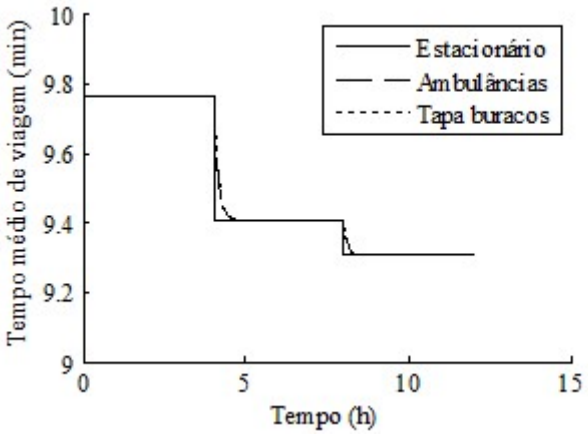
Grupo de servidores	Átomos		
	1	2	3
1	5	10	15
2	10	5	10
3	15	10	5

Os tempos médios de viagem para os chamados sujeitos à espera precisam de um ajuste com relação à aproximação utilizada no modelo hipercubo clássico, vista na Equação (5). Com as mudanças que as taxas de chegada sofrem ao longo do tempo, é necessário um tempo para que as proporções dos chamados na fila se alinhem às taxas atuais. Como os tempos de serviço são exponencialmente distribuídos, o tempo para o alinhamento é exponencial também. A Equação (60) mostra o processo para realizar esta aproximação. Na equação é feita o alinhamento é feito apenas a partir dos instantes t_i em que há alterações nas taxas de chegada ($\delta\lambda$). Os valores de $\bar{T}_{Q0}$ são obtidos pela Equação (5) do modelo hipercubo clássico.

$$\bar{T}_Q(t) = \bar{T}_{Q0}(t_i)\left(1 - \sum_{t_i:\delta\lambda} e^{-\mu(t)(t-t_i)}\right) + \sum_{t_i:\delta\lambda} \bar{T}_{Q0}(t_i)e^{-\mu(t)(t-t_i)}, \text{para } t_i < t \quad (60)$$

A Figura 47 ilustra esse efeito sobre o exemplo ilustrativo. Note que os modelos não-estacionários não possuem diferenças entre si nesta medida, visto que ela não leva em conta o vetor de probabilidade de estado em nenhum de seus termos.

Figura 47 – Tempo médio de viagem (minutos) para os chamados sujeitos à fila no exemplo ilustrativo do modelo hipercubo não-estacionário.

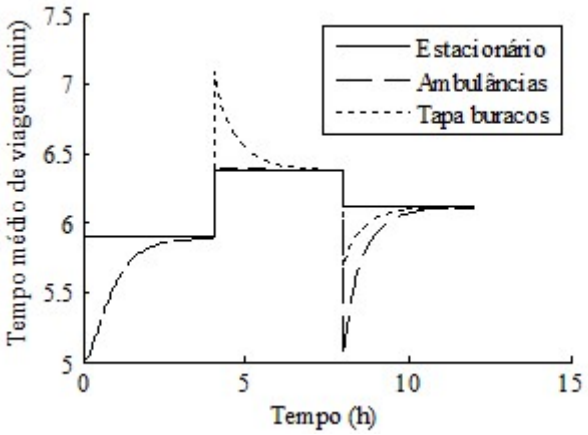


Os tempos médios de viagem dos sistemas são obtidos pela Equação (61). Essa equação é uma pequena adaptação da Equação (6) apenas considerando que os seus parâmetros estão em função do tempo.

$$\bar{T}(t) = \sum_{i=1}^{N} \sum_{j=1}^{N_A} f_{ij}^{(nq)}(t) t_{ij}(t) + P_Q'(t) \bar{T}_Q(t) \tag{61}$$

A Figura 48 mostra os tempos médios de viagem do sistema no exemplo ilustrativo. Os desvios com relação ao modelo estacionário podem variar entre 15 e -10% para o modelo com disciplina preemptiva. Para a disciplina exaustiva, os desvios relativos são majoritariamente positivos, chegando próximo à 17% no instante 8h.

Figura 48 – Tempo médio de viagem (minutos) do sistema para o exemplo ilustrativo do modelo hipercubo não-estacionário.

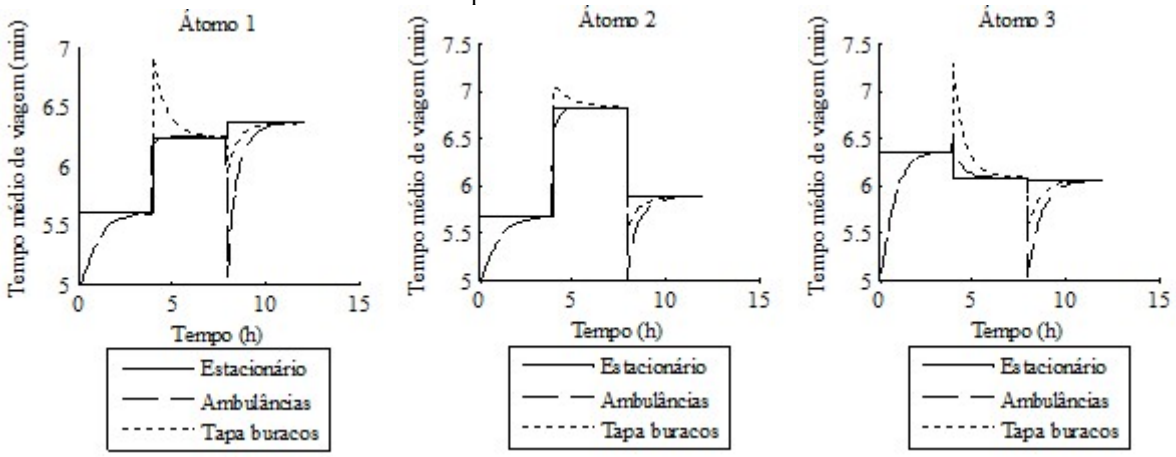


Os tempos médios de viagem aos átomos do sistema podem ser calculados pela Equação (62). Ela também é uma adaptação direta do modelo hipercubo clássico, da Equação (7).

$$\bar{T}_j(t) = \frac{\sum_{i=1}^{N} f_{ij}^{(nq)}(t) t_{ij}(t)}{\sum_{i=1}^{N} f_{ij}^{(nq)}(t)} \left(1 - P'_Q(t)\right) + \sum_{k=1}^{N_A} \left(\frac{\lambda_k(t)}{\lambda(t)}\right) \tau_{kj}(t) P'_Q(t) \qquad (62)$$

A Figura 49 mostra os tempos de viagem para cada um dos átomos do exemplo ilustrativo. Os desvios com relação ao modelo estacionário variam de 21% à -20% na disciplina preemptiva. Na disciplina exaustiva, os desvios variam de 21% até -7%, entre todos os átomos.

Figura 49 – Tempos médios de viagem (minutos) para os átomos para o exemplo ilustrativo do modelo hipercubo não-estacionário.

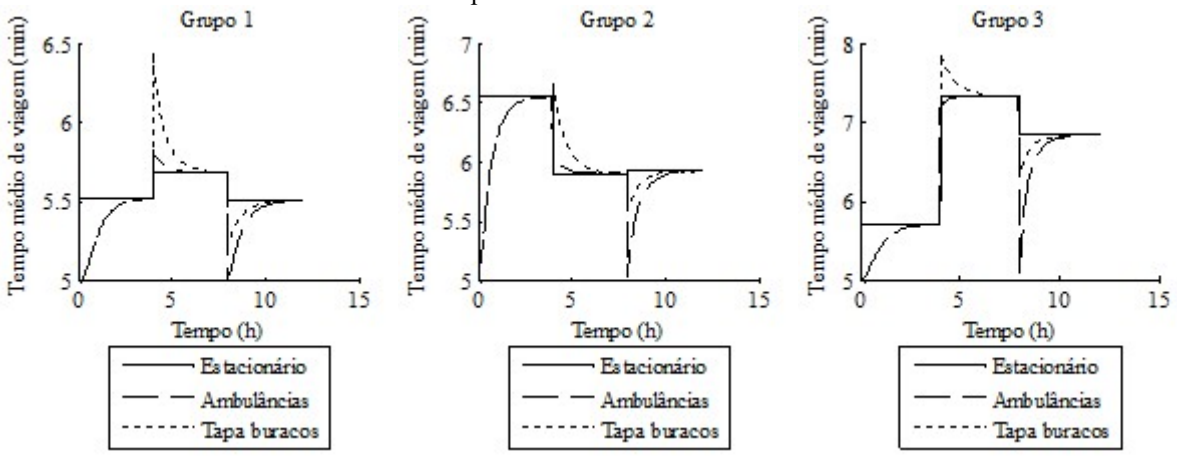


Por fim, os tempos médios de viagem dos servidores podem ser calculados pela Equação (63), adaptação direta da Equação (8) vista no modelo hipercubo clássico.

$$\overline{TU}_i(t) = \frac{\sum_{j=1}^{N_A} f_{ij}^{(nq)}(t) t_{ij}(t) + \frac{\mu_i(t)}{\mu(t)} \bar{T}_Q(t) P'_Q(t)}{\sum_{i=1}^{N} f_{ij}^{(nq)}(t) + \frac{\mu_i(t)}{\mu(t)} P'_Q(t)} \qquad (63)$$

A Figura 50 mostra os tempos médios de viagem dos grupos de servidores. Lembrando que os servidores de cada grupo são totalmente indistinguíveis. Os desvios com relação ao modelo estacionário variam de 24% até -13% na disciplina preemptiva. Na disciplina exaustiva, os desvios são de 26% até -4%.

Figura 50 – Tempos médios de viagem (minutos) dos grupos de servidores para o exemplo ilustrativo do modelo hipercubo não-estacionário.



Importante ressaltar que os momentos com maior diferença entre os tempos médios de viagem são os instantes após o início da operação e os instantes após as mudanças de turno. Esses são os instantes em que o sistema se encontra mais distante do equilíbrio.

5 APLICAÇÃO DO MODELO HIPERCUBO E EXTENSÕES NO SAMU-BAURU E ANÁLISE DOS RESULTADOS

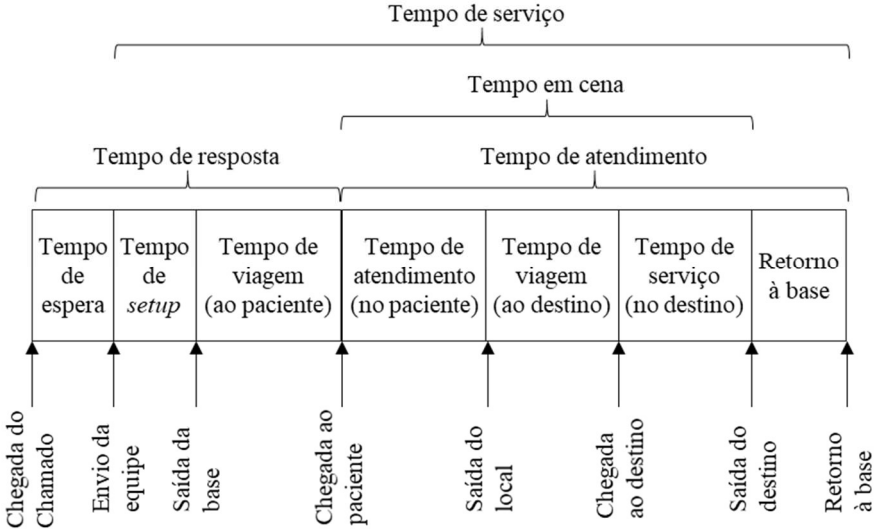
Este capítulo faz uma apresentação do SAMU de Bauru e segue com os resultados das aplicações dos modelos apresentados no Capítulo 4. Os resultados incluem os estudos do cenário original do sistema e avaliação de cenários alternativos. A primeira etapa envolve o estudo em período de pico e avaliação dos impactos de aumento na demanda e a inclusão de um novo servidor. A segunda etapa envolve o estudo do sistema ao longo de 24 horas, inicia-se com a validação das hipóteses, compara os resultados com modelos já existentes na literatura e, por fim, avalia cenários alternativos. Os cenários alternativos incluem a avaliação de aumento na demanda, em horários específicos do dia, e a alteração de pausas dos servidores.

5.1 O SAMU-Bauru

Desde 2010, o SAMU da região de Bauru integra 17 cidades numa parceria com prefeituras. A sede de atendimento está em Bauru, que deve receber os chamados e distribuí-los de acordo com a gravidade dos casos e de cada região. São aproximadamente 90 profissionais envolvidos. O serviço conta com 25 ambulâncias entre avançadas e básicas distribuídas entre os municípios cobertos. O Serviço regionalizado é responsável pela população de Bauru com aproximadamente 400 mil habitantes, e 800 mil pessoas na região como um todo. Esta análise foca apenas no SAMU da cidade de Bauru (JCNET, 2010).

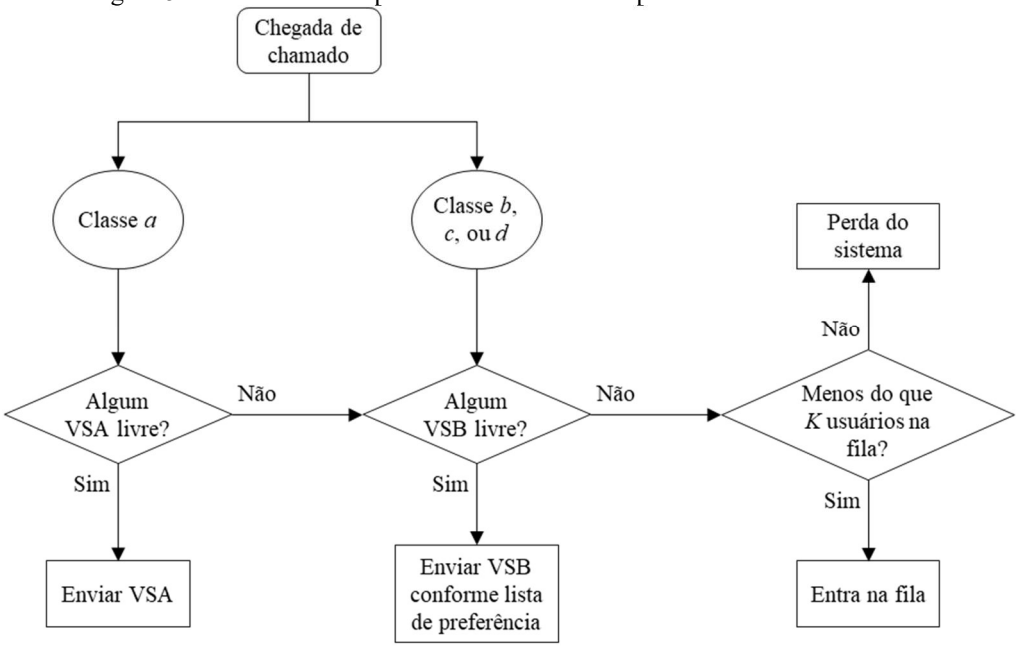
O processo de atendimento de um chamado em um SAMU pode ser descrito conforme a Figura 51. Para começar o processo é preciso que haja um chamado, normalmente realizada pelo telefone 192. Caso haja uma ambulância livre a equipe para o atendimento é enviada e a ambulância sai da base, assim que for feita a classificação (triagem) do chamado. Após isso, começa a viagem até o local do chamado. O intervalo de tempo entre o recebimento do chamado e a chegada ao local de atendimento representa o tempo de resposta. Começa então o atendimento ao usuário em si, o qual pode variar conforme os cuidados necessários. O paciente é então levado para um destino, o qual pode ser um hospital ou pronto-atendimento. Após deixar o paciente em seu destino, a ambulância retorna à base marcando o final do serviço (SCHMID, 2012).

Figura 51 – Linha do tempo com os eventos do atendimento de uma ambulância.



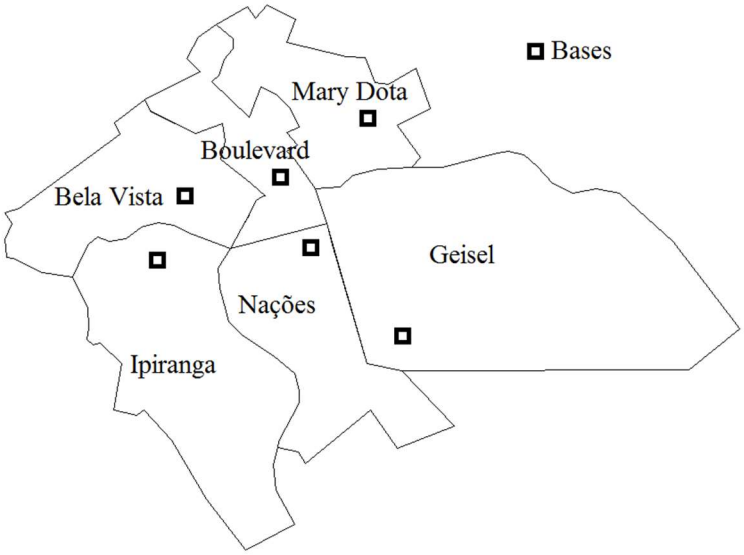
O serviço do SAMU envolve ainda a classificação dos chamados conforme sua gravidade. A classificação é feita utilizando um esquema de cores: vermelho (mais grave), amarelo, verde e azul (menos grave). O médico realiza a classificação dos chamados conforme as informações do solicitante no momento da ligação. Essa classificação também é utilizada para se decidir sobre a política de despacho das ambulâncias, já que os VSA's (Veículos de Serviço Avançado) são enviados apenas aos chamados classificados como vermelho, que incorrem risco de vida ao paciente. Em particular, o SAMU de Bauru tem dois servidores VSA´s, diferente do estudo de Souza *et al.* (2014) que tinha apenas um servidor preferencial no sistema estudado. A política de despacho do SAMU de Bauru pode ser vista como uma forma de reserva de capacidade dos VSA's. A Figura 52 ilustra o processo pelo qual é feito o despacho das ambulâncias. Também evidencia que os VSA's atendem apenas aos chamados vermelhos, chamados como classe *a*, enquanto os chamados amarelo, verde e azul são tratados por classes *b*, *c* e *d*, respectivamente.

Figura 52 – Política de despacho com reserva de capacidade do SAMU-Bauru.



As ambulâncias ficam distribuídas geograficamente, o sistema possui seis bases pela cidade onde as ambulâncias permanecem durante sua operação. Cada base é nomeada de acordo com o bairro em que estão localizadas. A Figura 53 mostra a localização das bases e as suas regiões de cobertura primária, são elas: Geisel (sede do sistema), Nações, Ipiranga, Mary Dota, Bela Vista e Boulevard. Cada base conta com ao menos uma ambulância VSB ao longo do dia. As exceções são a base Geisel concentra os dois VSA's e um VSB e a base Bela Vista com dois VSB's.

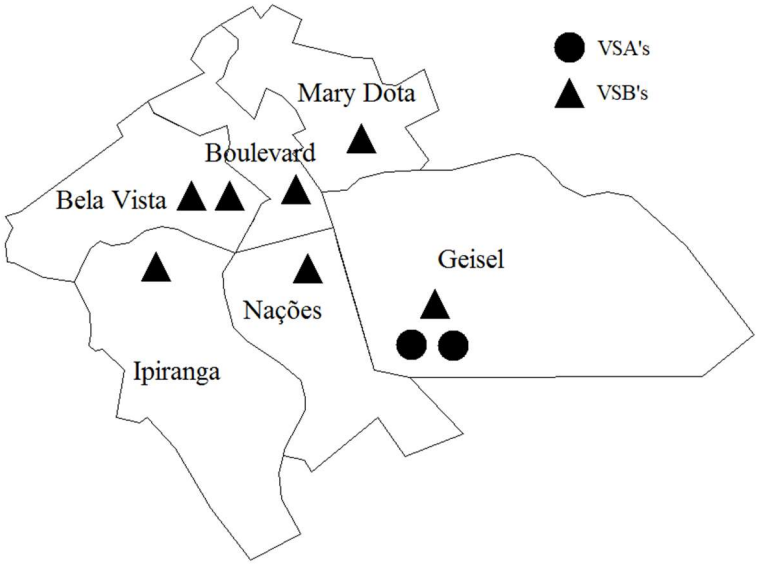
Figura 53 – Mapa do SAMU-Bauru com seus átomos e bases.



5.2 Resultados e análise do SAMU-Bauru no período de pico

Os resultados foram obtidos para o sistema SAMU-Bauru utilizando as extensões para o modelo estacionário da Seção 4.1. Os dados utilizados são referentes ao período entre às 12h e às 18h, considerado o período de pico do sistema. Este período possui a maior taxa de chegada de chamados com o menor desvio-padrão entre os 10 dias observados. Ghussn e Souza (2016) realizou a coleta dos dados e verificou as hipóteses de aplicação do modelo hipercubo clássico do SAMU-Bauru. É importante ressaltar que os dados foram coletados diretamente no SAMU-Bauru, observando os formulários de atendimento digitalizados um a um, pois o sistema SAMU-Bauru realiza o preenchimento manual (no computador) dos dados de cada atendimento, ainda não possui coleta automática dos dados por meio do uso de sistemas de GPS, nem o uso de um banco de dados detalhado. Nesse período, o sistema possui um total de nove ambulâncias em operação, distribuídas geograficamente pelas bases, conforme a Figura 54. Dessas, duas são VSA's (círculos) localizados na base Geisel. Os sete VSB's (triângulos) estão distribuídos ao longo das seis bases, sendo que a base Bela Vista possui dois VSB's.

Figura 54 – Mapa do SAMU-Bauru com a localização das ambulâncias durante o período de pico.



Os tempos de serviço tiveram sua hipótese de serem exponencialmente distribuídos rejeitada, contudo, optou-se por prosseguir com a modelagem e verificar os desvios em relação à amostra, como em Takeda *et al.* (2007) que não houve comprometimento da análise. As taxas de serviço dos servidores foram obtidas a partir dos seus respectivos tempos médios de serviço e podem ser vistas na Tabela 11. As duas ambulâncias avançadas, VSA's, localizadas na base

Geisel foram agrupadas em “GA”, assim como as duas ambulâncias básicas, VSB’s, localizadas na base Bela Vista em “BV”.

Tabela 11 – Taxas de serviço do SAMU-Bauru em seu período de pico.

Grupo de servidores	Tempo médio de serviço (min)	Taxa de serviço (μ)
Geisel Avançada (GA)	56,8	1,0562
Geisel Básica (GB)	47,2	1,2719
Nações (NÇ)	40,8	1,4720
Ipiranga (IP)	42,9	1,3998
Mary Dota (MD)	47,9	1,2517
Bela Vista (BV)	49,0	1,2254
Boulevard (BL)	45,8	1,3089

Os átomos do sistema foram enumerados para simplificar a separação dos subátomos: Geisel (1), Nações (2), Ipiranga (3), Mary Dota (4), Bela Vista (5) e Boulevard (6). Embora o SAMU-Bauru classifique seus chamados entre quatro classes de chamados, os chamados das classes *b*, *c* e *d* possuem mesma política de despacho. Por isso, optou-se por modelar o sistema utilizando apenas as camadas *a* (avançada, para chamados de classe *a*) e *b* (básica, para chamados de classe *b*, *c* ou *d*). As taxas de chegada foram calculadas a partir da proporção do número de chamados de cada subátomo em relação ao número total encontrado na amostra. Dessa maneira, a taxa total de chegada do sistema é distribuída proporcionalmente para cada subátomo, como mostra a Tabela 12.

Tabela 12 – Taxas de chegada do SAMU-Bauru em seu período de pico.

Subátomos	Nº de chamados	Proporção	Taxa de chegada (λ_{jk})
1 *a*	5	0,0248	0,0871
1 *b*	28	0,1386	0,4876
2 *a*	3	0,0149	0,0522
2 *b*	48	0,2376	0,8359
3 *a*	4	0,0198	0,0697
3 *b*	23	0,1139	0,4005
4 *a*	2	0,0099	0,0348
4 *b*	24	0,1188	0,4179
5 *a*	6	0,0297	0,1045
5 *b*	48	0,2376	0,8359
6 *a*	1	0,0050	0,0174
6 *b*	10	0,0495	0,1741
Total	202	1,0000	3,5176

O envio das ambulâncias é feito seguindo a matriz de preferência de despacho da Tabela 13, que traduz a política de despacho da Figura 52. As ambulâncias avançadas só podem atender aos chamados avançados, como forma de reserva de capacidade. Por esse motivo, a matriz de preferência não designa uma preferência para essas ambulâncias nos subátomos básicos. Além disso, as ambulâncias básicas são prioritárias para os chamados básicos locais. Quando ocupadas, outra ambulância básica é escolhida aleatoriamente entre as ambulâncias disponíveis mais próximas.

Tabela 13 – Matriz de preferência de despacho do SAMU-Bauru.

Subátomo	Grupo de servidores						
	GA	GB	NÇ	IP	MD	BV	BL
$1a$	1º	2º	4º	4º	3º	4º	4º
$1b$	-	1º	3º	3º	2º	3º	3º
$2a$	1º	3º	2º	4º	4º	4º	3º
$2b$	-	2º	1º	3º	3º	3º	2º
$3a$	1º	3º	3º	2º	3º	3º	3º
$3b$	-	2º	2º	1º	2º	2º	2º
$4a$	1º	4º	4º	4º	2º	4º	3º
$4b$	-	3º	3º	3º	1º	3º	2º
$5a$	1º	4º	4º	3º	4º	2º	3º
$5b$	-	3º	3º	2º	3º	1º	2º
$6a$	1º	4º	4º	4º	3º	4º	2º
$6b$	-	3º	3º	3º	2º	3º	1º

Os tempos médios de viagem entre os átomos da Tabela 14 foram estimados a partir da amostra. Nos casos em que não houve observação no período analisado, ou a amostra possuía menos do que 5 observações, utilizou-se a ferramenta Google Earth® para realizar a estimativa desses dados. Com a ferramenta é possível calcular as distâncias entre os centroides dos átomos, supondo uma velocidade de trânsito média de 40 km/h, estimou-se o tempo médio de viagem. Os dados indicados por asteriscos (*) são os obtidos por meio dessas estimativas.

Tabela 14 – Tempos médios de viagem entre os átomos do SAMU-Bauru.

τ_{ij}	$1a$	$1b$	$2a$	$2b$	$3a$	$3b$	$4a$	$4b$	$5a$	$5b$	$6a$	$6b$
$1a$	10	10	13,3	13,3	13,8	13,8	13,5	13,5	13,8*	13,8*	11	11
$1b$	10	10	13,3	13,3	13,8	13,8	13,5	13,5	13,8*	13,8*	11	11
$2a$	13,3	13,3	8,3	8,3	11,3	11,3	17*	17*	8,5	8,5	5,6	5,6
$2b$	13,3	13,3	8,3	8,3	11,3	11,3	17*	17*	8,5	8,5	5,6	5,6
$3a$	13,8	13,8	11,3	11,3	7,5	7,5	18*	18*	15,3	15,3	13*	13*

3*b*	13,8	13,8	11,3	11,3	7,5	7,5	18*	18*	15,3	15,3	13*	13*
4*a*	13,5	13,5	17*	17*	18*	18*	8,4	8,4	5	5	11	11
4*b*	13,5	13,5	17*	17*	18*	18*	8,4	8,4	5	5	11	11
5*a*	13,8*	13,8*	8,5	8,5	15,3	15,3	5	5	8,4	8,4	7,4	7,4
5*b*	13,8*	13,8*	8,5	8,5	15,3	15,3	5	5	8,4	8,4	7,4	7,4
6*a*	11	11	5,6	5,6	13*	13*	11	11	7,4	7,4	7*	7*
6*b*	11	11	5,6	5,6	13*	13*	11	11	7,4	7,4	7*	7*

5.2.1 Resultados do modelo original

O modelo foi implementado computacionalmente utilizando o software MATLAB® e executado em um computador com processador Intel Core i5-2400 de 3,10GHz, com 8GB de memória RAM DDR3 em 1333MHz, em um sistema operacional Windows 10 de 64bits.

O modelo foi aplicado ao SAMU-Bauru com a finalidade de realizar uma análise descritiva do sistema, do ponto de vista de suas medidas de desempenho em seu período de pico.

As equações de equilíbrio são determinadas semelhantemente aos exemplos ilustrativos na Seção 4.1, modelo considerando agregação de estados. Dado que o SAMU-Bauru possui 9 ambulâncias e 4 delas (2 VSA's da base Geisel e 2 VSB's da base Bela Vista) agrupadas em dois grupos de dois servidores, há $M = 3 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 3 \cdot 2 = 288$ estados possíveis para o sistema sem fila. Para a construção dos estados de fila escolheu-se um limite de 5 usuários. O SAMU-Bauru não possui um limite para fila, mas esse foi considerado aceitável, já que a probabilidade de perda é da ordem de 10^{-5}, como mostra a Tabela 15. A tabela também mostra as probabilidades de o sistema estar vazio e saturado. As equações da fila podem ser geradas, lembrando que os VSA's atendem apenas aos chamados avançados, com um total de $Q = 20 + 5 + 5 = 30$ estados, considerando os estados saturados e semi-saturados.

Tabela 15 – Medidas de probabilidade do SAMU-Bauru para o modelo estacionário.

Medida	Probabilidade
Vazio	0,06376
Saturação	0,01181
Perda	$3{,}61\times10^{-5}$

A Tabela 16 mostra os tempos médios de espera em fila do modelo, calculados utilizando a Equação (17). Ao contrário das outras medidas de desempenho do modelo, nenhuma comparação com a amostra foi feita para os tempos médios de espera. Isso se deve aos dados fornecidos pelo SAMU-Bauru, que não continham informações sobre os tempos de

triagem ou pré-atendimento dos chamados, tornando os tempos médios de espera muito maiores do que os obtidos pelo modelo, em torno 90% menores. Como era de se esperar, devido à reserva de capacidade, os chamados avançados possuem um tempo médio de espera muito menor do que os vistos nos básicos.

Tabela 16 – Tempos médios de espera (minutos) do SAMU-Baru para o modelo estacionário.

Prioridade	Tempo médio de espera (minutos)
a	0,3
b	7,6
Média	6,8

A Tabela 17 mostra as cargas de trabalho dos servidores, calculadas utilizando a Equação (36). Os desvios médios são de 5,73%, contudo duas ambulâncias apresentaram desvios relativos do modelo em relação à amostra superiores à 10%. A ambulância NÇ ficou 16% mais ocupada no modelo, enquanto a ambulância MD ficou 10% menos.

Tabela 17 – Cargas de trabalho dos servidores do SAMU-Bauru para o modelo estacionário.

Grupo de servidores	Carga de trabalho		
	Amostra	Modelo	Desvio relativo
GA	0,1657	0,1660	0,2%
GB	0,3800	0,3733	-1,8%
NÇ	0,3343	0,3888	16,3%
IP	0,2619	0,2852	8,9%
MD	0,3994	0,3575	-10,5%
BV	0,3672	0,3458	-5,8%
BL	0,3183	0,3250	2,1%

A Tabela 18 mostra os tempos médios de viagem dos servidores do SAMU-Bauru, calculados utilizando a Equação (21). Diferentemente das cargas de trabalho, os desvios relativos do modelo em relação à amostra para os tempos médios de viagem permaneceram sempre abaixo de 10%, sendo o maior desvio visto na ambulância IP, com 7,5%. Na média, os desvios relativos ficaram próximos à 5%.

Tabela 18 – Tempos médios de viagem dos servidores do SAMU-Bauru para o modelo estacionário.

Grupo de servidores	Tamanho da amostra	Tempo médio de viagem		
		Amostra	Modelo	Desvio relativo
GA	21	13,6	12,6	-6,9%
GB	29	10,8	11,1	3,6%

NÇ	21	8,4	8,7	3,6%
IP	22	8,8	9,4	7,5%
MD	27	10,9	10,4	-5,0%
BV	54	9,0	8,8	-2,0%
BL	25	7,3	7,9	7,3%

A Tabela 19 mostra os tempos médios de viagem aos átomos do SAMU-Bauru, calculados utilizando a Equação (7). Neste caso, apenas um átomo teve um desvio relativo do modelo em relação à amostra superior à 5%, no átomo Geisel. Na média, os desvios foram de 3,1%. O átomo Boulevard, mesmo com um número pequeno de chamados na amostra, com um total de 10 chamados, teve boa aderência ao modelo.

Tabela 19 – Tempos médios de viagem para os átomos do SAMU-Bauru para o modelo estacionário.

Átomo	Tamanho da amostra	Tempo médio de viagem		
		Amostra	Modelo	Desvio relativo
Geisel	33	10,5	11,1	5,3%
Nações	51	8,9	9,3	2,3%
Ipiranga	27	10,5	10,1	-3,8%
Mary Dota	22	9,8	10,0	-0,9%
Bela Vista	54	9,6	9,3	-2,6%
Boulevard	10	8,6	8,3	-4,0%

A Tabela 20 mostras os tempos médios de viagem para os subátomos do sistema, calculados utilizando a Equação (20). Os desvios relativos do modelo em relação à amostra ficaram pequenos, na média, para os chamados básicos. Os chamados avançados, que contam com amostras pequenas (menos do que 10 chamados para qualquer átomo), obtiveram desvios maiores, chegando à 27%.

Tabela 20 – Tempos médios de viagem aos subátomos do SAMU-Bauru para o modelo estacionário.

Subátomo	Tamanho da amostra	Tempo médio de viagem		
		Amostra	Modelo	Desvio relativo
1*a*	5	9,6	10,1	5,0%
1*b*	28	10,7	11,3	5,4%
2*a*	3	12,7	13,1	3,3%
2*b*	48	8,6	8,8	2,2%
3*a*	4	13,3	13,6	2,8%
3*b*	23	10,0	9,5	-5,3%
4*a*	2	10,5	13,3	26,8%
4*b*	20	9,8	9,4	-3,3%
5*a*	6	17,0	13,6	-20,3%

5*b*	48	8,7	8,8	1,7%
6*a*	1	14,0	10,9	-22,5%
6*b*	9	8,0	8,0	0,0%

Um dado importante da aderência do modelo à amostra neste estágio é que não foi necessário realizar a calibração dos tempos médios de serviço. Foi considerada uma tolerância de 0,1 chamados/hora para as taxas de serviço de cada servidor.

Mesmo com o sistema não tendo aderido totalmente às hipóteses do modelo, e devido à problemas com o tamanho das amostras, considerou-se a construção de cenários alternativos. Inicialmente, foram estudados os efeitos do aumento na demanda baseados em prospecções dos dados do SAMU-Bauru dos anos de 2012 e 2013 e aumentos atípicos. Depois avaliou-se a inclusão de uma nova ambulância para mitigar os efeitos no aumento na demanda prevista no médio e longo-prazos.

5.2.2 Cenários alternativos: aumento da demanda

O objetivo do estudo deste cenário é analisar o comportamento das medidas de desempenho do sistema sob dois pontos de vista. Primeiro, do ponto de vista do gestor do sistema, escolheu-se uma medida interna, a carga de trabalho. Segundo, do ponto de vista do usuário foram escolhidos os tempos médios de resposta aos átomos (tempo médio de viagem ao átomo + tempo médio de espera do sistema) e os tempos médios de espera dos chamados.

Para este cenário, encontrou-se a tendência do sistema para o ano seguinte aos anos dos dados coletados. Para tanto, foram utilizados três meios de cálculo para a previsão:

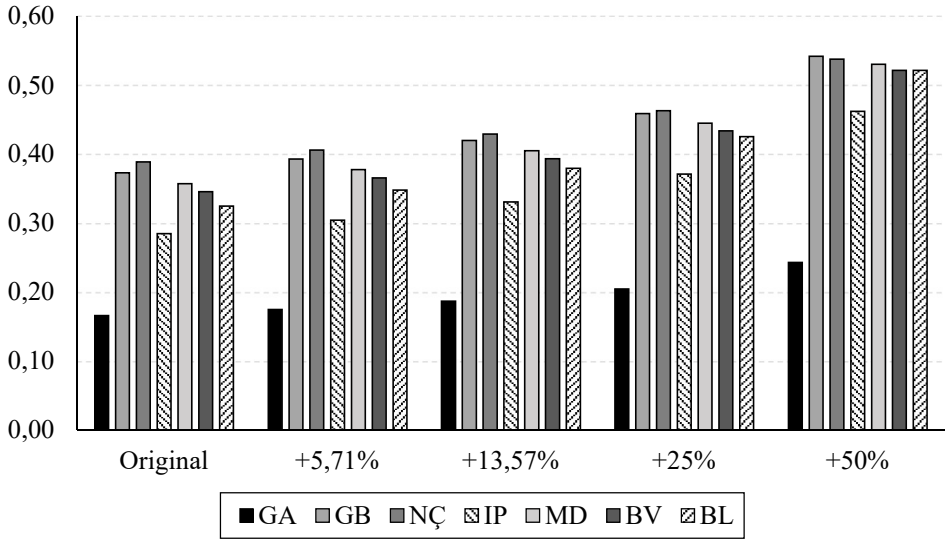
- Proporção dos chamados (13,57%): Encontrou-se o aumento da proporção dos chamados no período de pico de 2012 para 2013 e supôs-se que essa taxa de aumento se manteria para o ano seguinte igualmente para todos átomos;
- Análise de série temporal (5,71%): Através do resumo dos chamados mês a mês de 2012 e 2013, buscou-se encontrar previsões para o ano seguinte até o período de pico por meio do método da decomposição (encontrado no Anexo C), onde foi escolhido o mês de setembro, arbitrariamente, para estudo do aumento na demanda;
- Outras previsões (25% e 50%): escolha arbitrária de aumentos buscando entender situações atípicas extremas.

Os cenários de aumento na demanda mostram o impacto sobre o sistema causado pelo aumento no número médio de usuários que chegam no sistema, de forma a torná-lo mais

congestionado. Isso aumenta os tempos de espera em fila e aumenta o número de atendimentos realizados pelos *backups*.

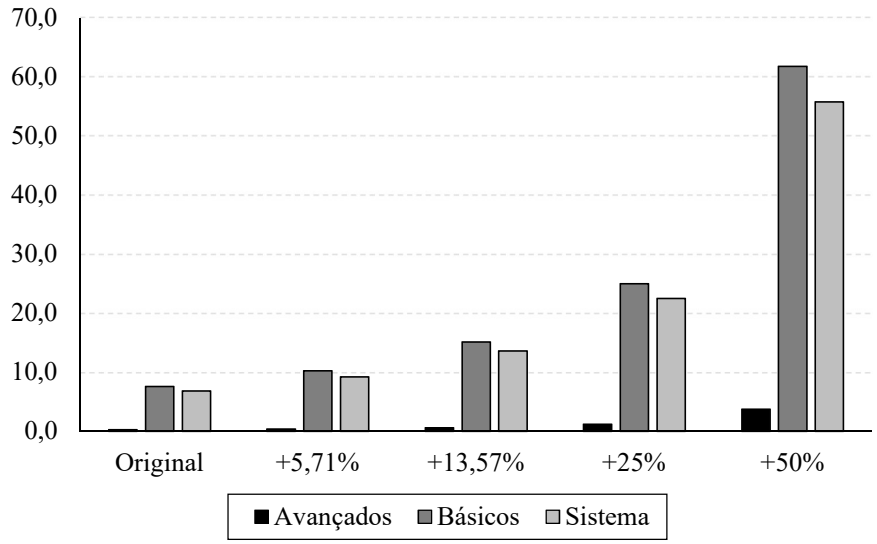
A Figura 55 coloca lado a lado as cargas de trabalho dos servidores em cada previsão de aumento na demanda. Embora todas ambulâncias tenham sido afetadas, duas ambulâncias foram mais afetadas pelos aumentos na demanda do que as demais. As ambulâncias IP e BL obtiveram um aumento na carga de trabalho superior ao aumento na demanda. No cenário com acréscimo de 5,71%, a ambulância BL teve um aumento de 7,1% em sua carga de trabalho, enquanto a ambulância IP, 6,7%. Para o cenário com acréscimo de 50%, ambas sofreram aumentos superiores à 60% em suas cargas de trabalho. Por outro lado, a ambulância menos afetada em todos cenários foi a NÇ, com aumentos inferiores ao aumento na demanda.

Figura 55 – Cargas de trabalho do SAMU-Bauru para o cenário de aumento na demanda no modelo estacionário.



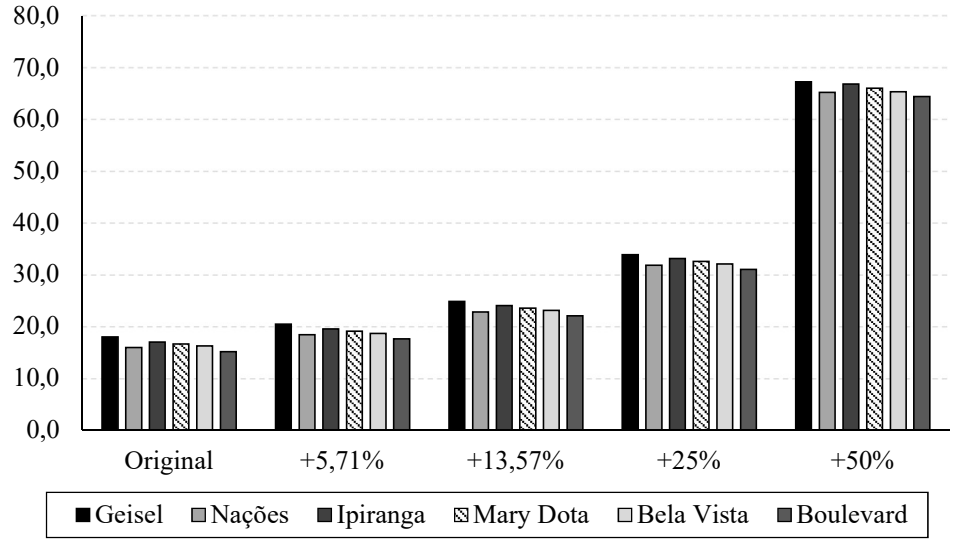
A Figura 56 mostra os tempos médios de espera para cada tipo de chamado nos cenários de aumento na demanda. Diferentemente das cargas de trabalho, os tempos médios de espera foram severamente impactados mesmo pelos menores aumentos na demanda. Por exemplo, o aumento de 13,57% na demanda já significa um aumento próximo à 100% nos tempos médios de espera do sistema. Sendo que os chamados avançados, embora continuem muito inferiores aos chamados básicos, são os mais afetados em termos relativos, com um aumento de 1300% em um aumento de 50% na demanda. Contudo os chamados avançados continuam com um tempo de espera pequeno, em torno de 3,8 minutos, em um aumento de 50% na demanda. Por outro lado, os chamados básicos sem encontram em uma situação preocupante com uma espera próxima a meia hora já com o aumento de 25% e superior a uma hora com o aumento de 50%.

Figura 56 – Tempos médios de espera em fila (minutos) do SAMU-Bauru para o cenário de aumento na demanda no modelo estacionário.



A Figura 57 mostra os tempos médios de resposta para os átomos nos cenários de aumento na demanda. Assim como os tempos médios de espera, o impacto é muito superior aos aumentos na demanda. Com um aumento de 25% na demanda, os tempos médios de resposta já crescem 97%. O átomo mais afetado pelo aumento na demanda, em termos relativos, foi o Boulevard, átomo que possui o menor tempo médio de resposta, cerca de 15 minutos no cenário original. O aumento na demanda não alterou a posição dos átomos, em termos de seus tempos médios de resposta. O átomo Geisel permaneceu com o maior tempo médio de resposta em todos cenários, enquanto Boulevard permaneceu com o menor.

Figura 57 – Tempos médios de resposta (minutos) aos átomos do SAMU-Bauru para o cenário de aumento na demanda no modelo estacionário.



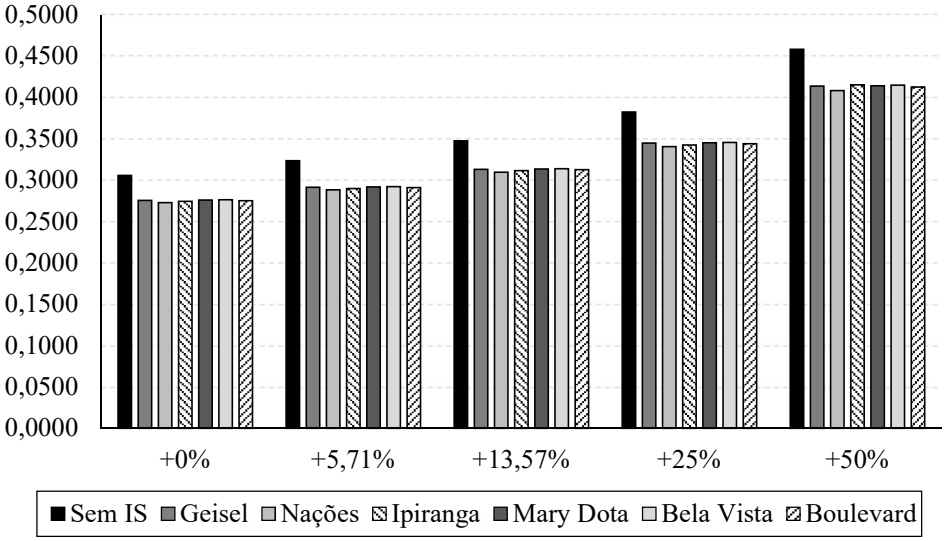
Esses resultados mostram a importância de se observar os dados tanto do ponto de vista do gestor do sistema, como do usuário. As medidas internas, do ponto de vista do gestor, apresentaram comportamento presumível. Por outro lado, nas medidas externas, do ponto de vista do usuário, os aumentos na demanda causam um impacto muito grande. Pode-se dizer que alguns resultados seriam alarmantes do ponto de vista dos usuários, quando se coloca em perspectiva que o sistema se trata de um SAE.

5.2.3 Cenários alternativos: inclusão de ambulância

Para cada um dos cenários de aumento na demanda foram analisados o impacto de se incluir uma nova ambulância (VSB) no sistema, essa análise é importante para permitir a mensuração do quanto e aonde a inclusão de um servidor pode ser melhor para o desempenho do sistema. A ambulância foi testada em cada uma das bases (átomos) já existentes do SAMU-Bauru. A nova ambulância utilizou os mesmos dados (tempos de atendimento e localização) que as ambulâncias já localizadas no átomo em que ela foi inserida. A avaliação da mitigação dos efeitos do aumento na demanda foi feita através de dois pontos de vista. Primeiro, do ponto de vista do gestor do sistema, através da carga de trabalho média do sistema. Segundo, do ponto de vista dos usuários, por meio do tempo médio de resposta do sistema. Também foram mensurados os desvios padrões das cargas de trabalho e dos tempos médios de resposta aos átomos como forma de verificar a situação mais equilibrada para o sistema.

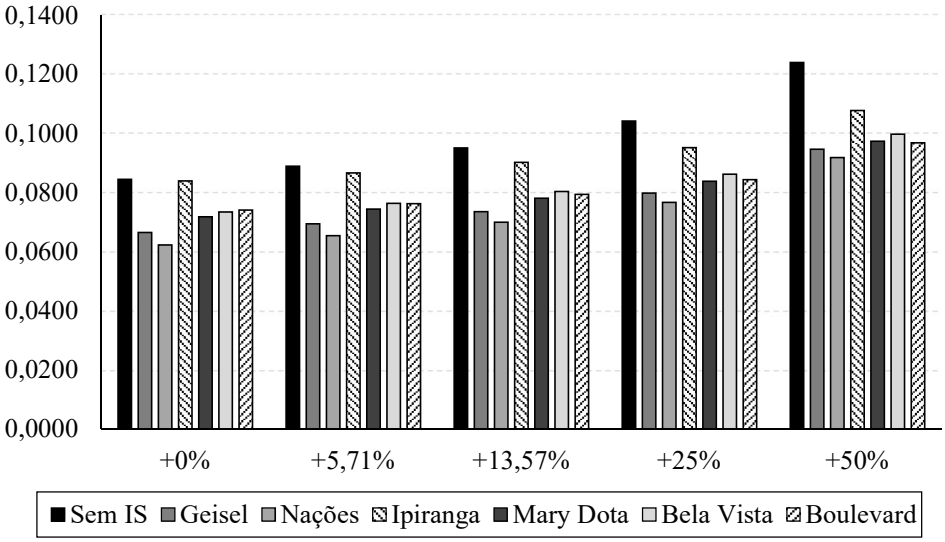
A Figura 58 mostra as médias das cargas de trabalho para cada localização da nova ambulância. A inclusão de uma nova ambulância consegue absorver o aumento na demanda até o cenário com 13,57% de acréscimo. Em todos cenários, a melhor localização para a nova ambulância foi o átomo Nações. Além disso, exceto no cenário com aumento de 50% na demanda, o átomo Bela Vista se mostrou a pior situação. Com o aumento de 50%, o átomo Ipiranga se mostrou a pior opção, com uma carga média de trabalho de 0,4153.

Figura 58 – Cargas médias de trabalho do SAMU-Bauru para os cenários com inclusão de nova ambulância no modelo estacionário.



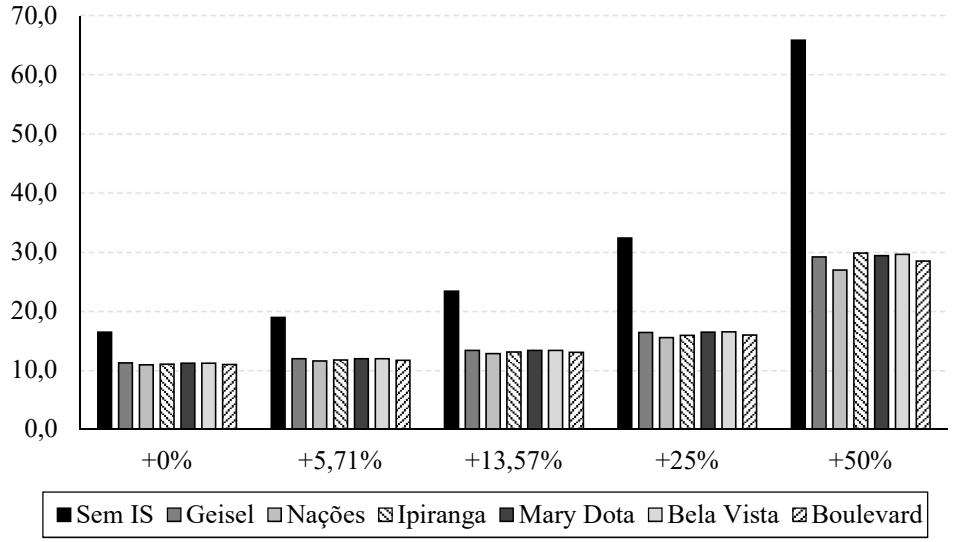
A Figura 59 mostra o desvio-padrão das cargas de trabalho para cada localização da nova ambulância. Em todos cenários, a localização que teve a maior desequilíbrio das cargas de trabalho foi o átomo Ipiranga. Sendo que, para os acréscimos de 5,71% e 13,57%, o desequilíbrio ficou bastante próximo ao cenário sem o novo servidor. A localização que obteve o melhor equilíbrio entre as cargas de trabalho foi o átomo Nações, para todos cenários de aumento na demanda.

Figura 59 – Desvio-padrão das cargas de trabalho do SAMU-Bauru para os cenários com inclusão de nova ambulância no modelo estacionário.



A Figura 60 mostra os tempos médios de resposta para o sistema para cada localização da nova ambulância. A inclusão de uma nova ambulância é capaz de absorver um acréscimo de 25% na demanda sem que o tempo médio de resposta do sistema aumente. Novamente, a localização com o melhor desempenho é o átomo Nações. Mesmo com um aumento de 25% na demanda, os tempos médios de resposta ficam em torno de 15,5 minutos, 1 minuto a menos que o observado no cenário original. Por outro lado, com um acréscimo de até 25% a pior localização é o átomo Bela Vista. Para um acréscimo de 50% a pior localização está no átomo Ipiranga.

Figura 60 – Tempos médios de resposta (minutos) do sistema do SAMU-Bauru para os cenários com inclusão de nova ambulância no modelo estacionário.



Outro resultado interessante pode ser observado no cenário com aumento de 50% na demanda, com a nova ambulância localizada no átomo Ipiranga. Esse é o único cenário em que foi necessário realizar uma iteração da calibragem dos tempos médios de serviço dentre todos os cenários estudados. Além disso, os tempos computacionais ficaram entre 1 e 2 segundos em todos cenários.

De maneira geral, a melhor localização para a nova ambulância foi o átomo Nações. Esse resultado pode ser observado tanto do ponto de vista do gestor, onde houve a maior redução nas cargas de trabalho e onde as cargas de trabalho ficaram mais equilibradas. Além disso, do ponto de vista do usuário, o átomo Nações também foi o melhor, obtendo o menor tempo de resposta para todos os cenários. Com isso, a possível situação alarmante causada pelo aumento na demanda pode ser mitigada com sucesso.

5.3 Resultados e análise do SAMU-Bauru com parâmetros variando no tempo

Esta seção apresenta os resultados para a etapa do estudo do SAMU-Bauru que considera parâmetros variáveis no tempo. Primeiro, faz-se uma verificação das hipóteses para aplicação. Depois faz-se um estudo da situação original do sistema, comparando os resultados do modelo não-estacionário a uma aproximação estacionária. Finalmente apresenta-se a avaliação de cenários alternativos como o aumento na demanda em alguns horários do dia e a avaliação da troca de horário para pausas dos servidores.

5.3.1 Verificação das hipóteses do modelo hipercubo não-estacionário no SAMU-Bauru

Esta seção traz as hipóteses levantadas ao longo da Seção 4.2.3. Elas são verificadas para a aplicação do modelo hipercubo não-estacionário ao SAMU-Bauru.

5.3.1.1 Existência de átomos geográficos

Como mostrado na Figura 53, o SAMU-Bauru tem seu atendimento distribuído ao longo de seis átomos. Cada átomo possui uma base com ambulâncias em pelo menos uma parte do dia. A identificação dos átomos e seus tamanhos permanecem inalterados ao longo da operação do sistema. Os átomos foram separados em camadas, pelo processo de *layering*, de maneira a trabalhar com duas classificações dos chamados: Avançados e Básicos. Com isso, o sistema é composto por 12 subátomos ao todo.

5.3.1.2 Processo de chegada

Os dados foram retirados dos relatórios de atendimento, mesmos vistos em Ghussn e Souza (2016), para 10 dias do mês de setembro do ano de 2013. Para a análise os chamados foram separados em intervalos de uma hora, $[t, t+1)$. A Tabela 21 mostra a quantidade de chamados para cada subátomo nesses intervalos de hora.

Tabela 21 – Número de chamados para cada subátomo em cada hora de operação do SAMU-Bauru.

Horário	Subátomos											
	$1a$	$1b$	$2a$	$2b$	$3a$	$3b$	$4a$	$4b$	$5a$	$5b$	$6a$	$6b$
0-1h	1	1	0	4	0	3	0	4	0	3	0	3
1-2h	0	1	0	1	0	4	0	2	1	4	0	0
2-3h	0	2	0	1	0	2	0	1	0	3	0	1
3-4h	0	2	0	1	0	2	0	3	0	5	0	0
4-5h	0	1	0	0	0	1	0	6	0	0	0	0

5-6h	2	1	1	1	0	3	0	1	0	5	0	0
6-7h	0	2	0	1	0	1	0	1	0	1	0	0
7-8h	0	2	0	7	0	6	0	2	2	10	0	5
8-9h	2	6	0	5	1	5	0	6	0	6	0	1
9-10h	1	5	1	7	0	8	0	3	0	4	0	5
10-11h	1	6	1	7	0	7	0	4	0	8	0	1
11-12h	1	7	0	7	2	5	0	2	1	9	0	1
12-13h	0	5	0	9	1	4	0	3	1	11	0	0
13-14h	1	2	0	7	3	6	0	7	1	13	0	4
14-15h	1	7	1	9	1	4	1	3	3	6	0	6
15-16h	1	7	0	9	1	9	0	3	0	8	0	0
16-17h	4	4	2	6	1	4	1	6	0	6	0	3
17-18h	0	6	0	9	1	4	0	7	2	7	0	4
18-19h	2	6	0	3	1	7	0	5	0	8	0	1
19-20h	1	1	0	6	0	3	0	3	0	7	1	3
20-21h	1	3	0	7	1	5	0	4	2	4	0	2
21-22h	1	7	1	2	1	5	1	1	0	4	0	1
22-23h	1	2	0	5	0	6	0	1	0	8	1	3
23-24h	1	4	1	1	2	3	0	2	0	2	0	0

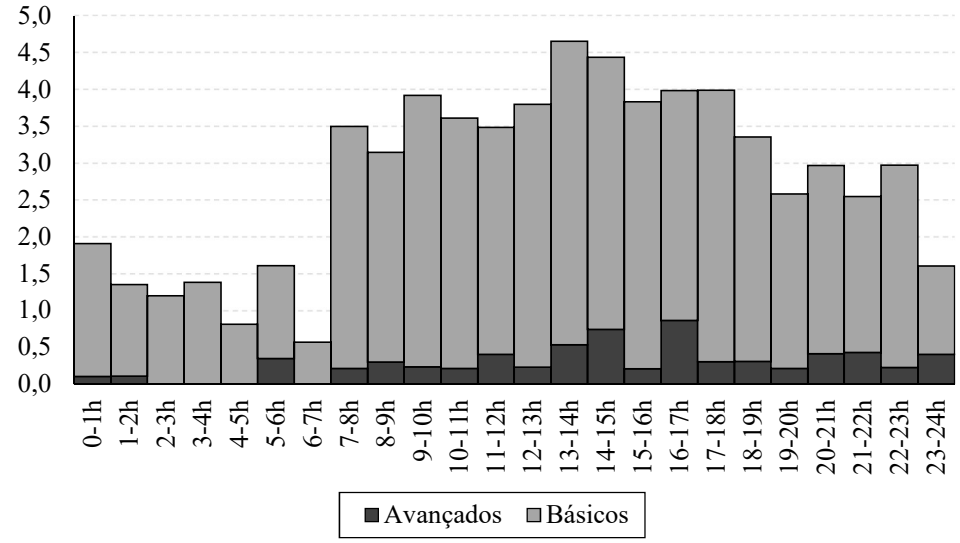
É necessário realizar testes de aderência aos dados a fim de verificar estatisticamente a hipótese de o processo de chegada corresponder a um processo de Poisson não-homogêneo. Os métodos utilizados são baseados em duas transformações dos dados para o teste de Kolmogorov-Smirnov. A primeira transformação é baseada na propriedade condicional uniforme do processo de Poisson, enquanto a segunda é feita sobre o trabalho de Lewis (1965); ver Anexo D. Para a propriedade condicional uniforme do processo de Poisson, os horários das chegadas, quando ordenadas, são uniformemente distribuídos. A transformação realizada no trabalho de Lewis (1965) faz com que os dados transformados, quando ordenados, também são uniformemente distribuídos entre 0 e 1. Considera-se que a taxa de chegada varia suavemente, tornando possível verificar que, para intervalos de tempo relativamente pequenos, a taxa de chegada seja aproximadamente constante. Com isso, utiliza-se as transformações citadas para verificar se o processo de chegada no dado intervalo de tempo é aproximadamente constante. A Tabela 22 resume os resultados dos testes. Lembrando que os dados foram desarredondados sem prejuízo da análise, como mostra Kim e Whitt (2014). Note que como nenhum intervalo de tempo foi rejeitado em ambos testes, considera-se a hipótese não rejeitada. Embora seja importante ressaltar, como vista na Tabela 21, que as amostras para vários intervalos de tempo são pequenas e, por isso, a validade do teste de hipótese pode ser questionável.

Tabela 22 – Resumo dos testes de hipótese para as taxas de chegada ao longo do tempo.

Horário	CU-KS				Lewis KS			
	Arredondado		Desarredondado		Arredondado		Desarredondado	
	H_0	valor-p	H_0	valor-p	H_0	valor-p	H_0	valor-p
0-1h	Ok	0,7511	Ok	0,7489	Ok	0,3699	Ok	0,3264
1-2h	Não	0,0082	Não	0,0082	Ok	0,1560	Ok	0,1561
2-3h	Ok	0,3005	Ok	0,3058	Ok	0,2823	Ok	0,3057
3-4h	Ok	0,9989	Ok	0,9991	Não	0,0092	Não	0,0092
4-5h	Ok	0,9500	Ok	0,9489	Ok	0,1724	Ok	0,1809
5-6h	Ok	0,5206	Ok	0,5305	Ok	0,5034	Ok	0,5277
6-7h	Ok	0,6054	Ok	0,6123	Ok	0,3348	Ok	0,3368
7-8h	Ok	0,2904	Ok	0,2938	Ok	0,6666	Ok	0,6012
8-9h	Ok	0,6937	Ok	0,6970	Ok	0,7292	Ok	0,7521
9-10h	Não	0,0215	Não	0,0226	Ok	0,2113	Ok	0,1912
10-11h	Ok	0,9644	Ok	0,9616	Ok	0,0677	Ok	0,0749
11-12h	Ok	0,1970	Ok	0,1979	Ok	0,4860	Ok	0,5301
12-13h	Ok	0,9632	Ok	0,9591	Ok	0,9536	Ok	0,8273
13-14h	Ok	0,5366	Ok	0,5498	Não	0,0496	Ok	0,1375
14-15h	Ok	0,5111	Ok	0,5019	Ok	0,7750	Ok	0,8011
15-16h	Não	0,0452	Não	0,0448	Ok	0,5785	Ok	0,5463
16-17h	Ok	0,0639	Ok	0,0674	Ok	0,3746	Ok	0,4140
17-18h	Ok	0,0596	Ok	0,0589	Ok	0,4858	Ok	0,5261
18-19h	Ok	0,5327	Ok	0,5387	Não	0,0251	Não	0,0346
19-20h	Ok	0,4945	Ok	0,4975	Ok	0,9099	Ok	0,8322
20-21h	Ok	0,3538	Ok	0,3547	Ok	0,7174	Ok	0,7765
21-22h	Ok	0,9568	Ok	0,9585	Ok	0,0901	Ok	0,0757
22-23h	Ok	0,3908	Ok	0,4029	Ok	0,5543	Ok	0,5818
23-24h	Ok	0,2956	Ok	0,2992	Ok	0,2415	Ok	0,2401

A Figura 61 mostra a evolução das taxas de chegada do sistema ao longo do tempo. Os chamados foram separados apenas em Avançados e Básicos, sem distinção dos átomos de origem. Lembrando que as taxas foram consideradas aproximadamente constantes nos intervalos $[t, t+1)$. É possível observar que a partir das 7 horas o número de chamados cresce fortemente, tendo seu pico entre às 13 e 14 horas. A partir disso, as taxas diminuem até as 18 horas e permanecem aproximadamente estáveis até às 23 horas, voltando a diminuir durante a madrugada até o amanhecer às 7 horas.

Figura 61 – Resumo das taxas de chegada ao longo das 24 horas do SAMU-Bauru.



A Tabela 23 mostra as taxas de chegada para cada subátomo. As taxas foram obtidas a partir da taxa total do sistema, no dado intervalo de tempo, que foi distribuída proporcionalmente ao número de chamados para cada subátomo. A taxa total de chegada do sistema, por sua vez, foi obtida a partir da média dos intervalos de tempo entre as chegadas do sistema. Os subátomos em que não foram observados nenhum chamado na amostra, ao invés de possuírem uma taxa de chegada nula, considerou-se uma taxa de 0,0001 chamados/hora.

Tabela 23 – Taxas de chegada ao longo das 24 horas do SAMU-Bauru.

Horário	Subátomos											
	1 *a*	1 *b*	2 *a*	2 *b*	3 *a*	3 *b*	4 *a*	4 *b*	5 *a*	5 *b*	6 *a*	6 *b*
0-1h	0,1003	0,1003	0,0001	0,4010	0,0001	0,3008	0,0001	0,4010	0,0001	0,3008	0,0001	0,3008
1-2h	0,0001	0,1037	0,0001	0,1037	0,0001	0,4149	0,0001	0,2074	0,1037	0,4149	0,0001	0,0001
2-3h	0,0001	0,2389	0,0001	0,1195	0,0001	0,2389	0,0001	0,1195	0,0001	0,3584	0,0001	0,1195
3-4h	0,0001	0,2118	0,0001	0,1059	0,0001	0,2118	0,0001	0,3177	0,0001	0,5295	0,0001	0,0001
4-5h	0,0001	0,1010	0,0001	0,0001	0,0001	0,1010	0,0001	0,6058	0,0001	0,0001	0,0001	0,0001
5-6h	0,2293	0,1146	0,1146	0,1146	0,0001	0,3439	0,0001	0,1146	0,0001	0,5732	0,0001	0,0001
6-7h	0,0001	0,1887	0,0001	0,0943	0,0001	0,0943	0,0001	0,0943	0,0001	0,0943	0,0001	0,0001
7-8h	0,0001	0,2058	0,0001	0,7202	0,0001	0,6173	0,0001	0,2058	0,2058	1,0289	0,0001	0,5144
8-9h	0,1967	0,5901	0,0001	0,4918	0,0984	0,4918	0,0001	0,5901	0,0001	0,5901	0,0001	0,0984
9-10h	0,1153	0,5766	0,1153	0,8072	0,0001	0,9225	0,0001	0,3460	0,0001	0,4613	0,0001	0,5766
10-11h	0,1032	0,6190	0,1032	0,7221	0,0001	0,7221	0,0001	0,4126	0,0001	0,8253	0,0001	0,1032
11-12h	0,0996	0,6974	0,0001	0,6974	0,1993	0,4982	0,0001	0,1993	0,0996	0,8967	0,0001	0,0996
12-13h	0,0001	0,5589	0,0001	1,0060	0,1118	0,4471	0,0001	0,3353	0,1118	1,2295	0,0001	0,0001
13-14h	0,1058	0,2117	0,0001	0,7409	0,3175	0,6351	0,0001	0,7409	0,1058	1,3759	0,0001	0,4234
14-15h	0,1057	0,7401	0,1057	0,9515	0,1057	0,4229	0,1057	0,3172	0,3172	0,6343	0,0001	0,6343
15-16h	0,1009	0,7063	0,0001	0,9081	0,1009	0,9081	0,0001	0,3027	0,0001	0,8072	0,0001	0,0001
16-17h	0,4308	0,4308	0,2154	0,6463	0,1077	0,4308	0,1077	0,6463	0,0001	0,6463	0,0001	0,3231
17-18h	0,0001	0,5990	0,0001	0,8985	0,0998	0,3993	0,0001	0,6988	0,1997	0,6988	0,0001	0,3993
18-19h	0,2034	0,6103	0,0001	0,3051	0,1017	0,7120	0,0001	0,5086	0,0001	0,8137	0,0001	0,1017
19-20h	0,1032	0,1032	0,0001	0,6194	0,0001	0,3097	0,0001	0,3097	0,0001	0,7226	0,1032	0,3097
20-21h	0,1024	0,3071	0,0001	0,7165	0,1024	0,5118	0,0001	0,4094	0,2047	0,4094	0,0001	0,2047
21-22h	0,1061	0,7426	0,1061	0,2122	0,1061	0,5304	0,1061	0,1061	0,0001	0,4244	0,0001	0,1061
22-23h	0,1101	0,2201	0,0001	0,5503	0,0001	0,6603	0,0001	0,1101	0,0001	0,8804	0,1101	0,3302

23-24h	0,1001	0,4004	0,1001	0,1001	0,2002	0,3003	0,0001	0,2002	0,0001	0,2002	0,0001	0,0001

5.3.1.3 Tempos médios de viagem

Os tempos médios de viagem entre os átomos do sistema foram considerados constantes ao longo das 24 horas de operação do SAMU-Bauru. Por isso, a matriz de τ_{ij} utilizada é a mesma presente na Tabela 14 com suas estimativas.

5.3.1.4 Servidores

As ambulâncias do SAMU-Bauru operam 24 horas em todas as bases do sistema. A configuração é a mesma durante todo o período de análise. Sendo que o sistema conta, ao todo, com 2 VSA's e 7 VSB's.

As equipes operam em dois turnos de 12 horas que se iniciam às 7 e às 19 horas. Durante a mudança de turno as equipes são trocadas por completo. Caso uma equipe esteja operando no momento da troca de turno, esta finaliza o atendimento antes de retornar à base, caracterizando uma disciplina exaustiva de fim de turno. No entanto, devido à limitação no tamanho da frota de ambulâncias, as equipes que iniciam um turno precisam esperar as equipes anteriores finalizarem seus atendimentos para que possam começar a operar.

Cada turno possui 1 hora para realizar suas refeições. Contudo, as refeições não possuem horário determinado para começar. O SAMU-Bauru também não guarda registros sobre os horários de início das refeições, apenas realiza um controle diário para que o tempo de almoço não ultrapasse 1 hora. As refeições não podem ser interrompidas devido à forma de contrato dos funcionários das equipes. Segundo relato dos gestores do SAMU-Bauru, as equipes, cujo turno se inicia às 7 horas, costumam começar o horário de almoço em um período entre às 11 e às 13 horas. Enquanto as equipes cujo turno se inicia às 19 horas costuma começar o horário de jantar em um período entre às 20 e às 22 horas. Outras refeições podem ser realizadas ao longo do dia, porém estas podem ser interrompidas por força da chegada de uma nova ocorrência.

É importante ressaltar que as equipes não podem parar todas juntas para as refeições. Nos casos das bases com 2 ambulâncias, enquanto uma equipe tem o horário da refeição a outra continua operando. Nos casos das bases com apenas 1 ambulância, deve haver ao menos uma das equipes mais próximas operando durante o horário das refeições.

Dessa maneira, a modelagem dos servidores seguirá os seguintes parâmetros:

- As ambulâncias são modeladas como se não houvesse mudança de turno, já que não param de operar, como sugerido por Ingolfsson (2005) para a operação em sistemas com recursos escassos;
- Os horários das refeições foram modelados da seguinte maneira: gera-se um conjunto de horários das refeições aleatório, cujos inícios ocorrem entre 11 e 13 horas (para o almoço) e 20 e 22 horas (para o jantar), de maneira que todas equipes tenham suas refeições, sem interromper o atendimento por completo, mantendo ao menos o primeiro *backup* de cada átomo operando, ou uma das ambulâncias do grupo (caso dos VSB's do átomo Bela Vista e dos VSA's) e, por fim, o resultado final é dado pela média dos resultados dos conjunto de horários gerado.

5.3.1.5 Localização dos servidores

As ambulâncias possuem localização fixa. A Figura 54 mostra a localização das ambulâncias em suas bases distribuídas pela cidade. Em resumo, ao longo das 24 horas, a base Geisel conta com 2 VSA's e 1 VSB, as bases Nações, Ipiranga, Boulevard e Mary Dota contam com 1 VSB cada e a base Bela Vista conta com 2 VSB's, assim, não há mudanças de bases ao longo das 24 horas.

5.3.1.6 Despacho dos servidores

O SAMU-Bauru adota a política de enviar apenas uma ambulância para atender a um chamado. Na amostra, há apenas um caso em que duas ambulâncias foram enviadas, este caso foi excluído da amostra. Casos de grandes acidentes ou catástrofes envolvem outros sistemas de apoio como a defesa civil, os bombeiros e a polícia. A política do SAMU-Bauru ainda permite a formação de fila de espera. O sistema não possui um limite para a formação de fila predeterminado.

Para o modelo, considerou-se um limite de 5 usuários em fila. Esse número foi considerado adequado visto que o pico da probabilidade de haver mais usuários no sistema foi da ordem de 10^{-5}.

5.3.1.7 Política de despacho

A Tabela 24 mostra a matriz de preferência de despacho, que permanece inalterada ao longo das 24 horas de operação, para o SAMU-Bauru. A matriz é baseada no esquema visto na

Figura 52. A diferença é que aqui os VSA's podem atender aos chamados básicos, todavia são as últimas opções para tal. Isso é uma relaxação do esquema necessária, visto que o modelo hipercubo não-estacionário não abrange a reserva total de capacidade, ou o *backup* parcial.

Tabela 24 – Matriz de preferência do SAMU-Bauru ao longo do dia.

Subátomo	Grupo de servidores						
	GA	GB	NÇ	IP	MD	BV	BL
1 a	1º	2º	4º	4º	3º	4º	4º
1 b	4º	1º	3º	3º	2º	3º	3º
2 a	1º	3º	2º	4º	4º	4º	3º
2 b	4º	2º	1º	3º	3º	3º	2º
3 a	1º	3º	3º	2º	3º	3º	3º
3 b	4º	2º	2º	1º	2º	2º	2º
4 a	1º	4º	4º	4º	2º	4º	3º
4 b	4º	3º	3º	3º	1º	3º	2º
5 a	1º	4º	4º	3º	4º	2º	3º
5 b	4º	3º	3º	2º	3º	1º	2º
6 a	1º	4º	4º	4º	3º	4º	2º
6 b	4º	3º	3º	3º	2º	3º	1º

5.3.1.8 Tempos de serviço

Os tempos de serviço foram obtidos através dos relatórios de atendimento. O tempo de serviço foi definido como o intervalo de tempo entre o envio da equipe até o retorno à base, como visto na Figura 51. Os tempos de serviço foram divididos entre os dois turnos. Os testes foram realizados considerando que as taxas de serviço são constantes ao longo de cada turno.

Primeiramente se verificou a hipótese de que os tempos de serviço de cada grupo de ambulâncias em cada turno são exponencialmente distribuídos. Para um nível de significância $\alpha = 0{,}05$ apenas os VSA's do turno com início às 19 horas não puderam ter a hipótese nula utilizando o teste Kolmogorov-Smirnov.

Depois foi verificada a hipótese de os tempos de serviço variarem entre os turnos. Utilizando um teste ANOVA com nível de significância $\alpha = 0{,}05$. Apenas os servidores da base Bela Vista e da base Ipiranga tiveram diferenças significativas entre os turnos.

Tendo em vista estes resultados, optou-se por utilizar os mesmos tempos de serviço utilizados para o modelo estacionário da Seção 5.2, vistos na Tabela 11. Lembrando que ao longo dos períodos das refeições algumas ambulâncias não operam, portanto, são desconsideradas, juntamente de suas taxas de serviço, do modelo.

5.3.1.9 Dependência dos tempos de viagem

Assim como os tempos de serviço, os tempos de viagem foram comparados entre os dois turnos. Considerando um nível de significância $\alpha = 0{,}05$, apenas os servidores básicos das bases Geisel e Boulevard tiveram diferenças significativas entre os turnos. Por isso, optou-se por utilizar os mesmos tempos de viagem vistos no modelo estacionário da Seção 5.2, vistos na Tabela 14.

A Tabela 25 mostra que os tempos médios de viagem são uma variável de segunda ordem no tempo médio de serviço, já que não ultrapassam 25% do tempo médio de serviço, de forma de as variações dos tempos médios de viagem também foram consideradas secundárias. Em Takeda *et al.* (2007) é possível observar a mesma ordem de grandeza na verificação dessa relação.

Tabela 25 – Relação entre o tempo médio de viagem e o tempo médio de serviço para o SAMU-Bauru.

Ambulância	Tempo médio de viagem	Tempo médio de serviço	Relação
GA	13,6	56,8	0,24
GB	10,8	47,2	0,23
NÇ	8,4	40,8	0,21
IP	8,8	42,9	0,20
MD	10,0	47,9	0,21
BV	9,0	49,0	0,18
BL	7,3	45,8	0,16

É importante ressaltar que ainda não há um método para calibração dos tempos médios de serviço para o modelo hipercubo não-estacionário. Isto não foi considerado uma limitação forte para aplicação do modelo visto que os tempos médios de serviço foram calibrados em apenas uma situação durante o modelo estacionário.

5.3.1.10 Solução Inicial

Para estimar uma solução inicial para o modelo, realizou-se um período de aquecimento para o modelo. Esse período de aquecimento consiste em rodar o modelo uma única vez utilizando como solução inicial o sistema vazio, $P_{\{0000000\}}(0) = 1$. O último vetor de probabilidade encontrado, $P(24)$, foi utilizado como solução inicial, $P(0)$, do modelo definitivo.

Caso o sistema sofresse alguma alteração de final de turno em $t = 24$ horas, o vetor de probabilidade precisaria passar por uma transição instantânea antes de ser considerado como a solução inicial. Esta transição instantânea foi apresentada na Seção 4.2.1 para sistemas com disciplina exaustiva, como o SAMU-Bauru, e na Seção 4.2.2 para sistemas com disciplina preemptiva. Isto não foi necessário, visto que as trocas de turno ocorrem às 7h e às 19h e as refeições ocorrem entre às 11h e às 14h para o almoço e entre às 20h e às 23h para o jantar (considerando os horários de início e final das refeições).

5.3.2 Resultados do modelo original

O modelo foi implementado computacionalmente utilizando o software MATLAB®, utilizando a função "ode45" para resolver numericamente o sistema de equações diferenciais. Esta função utiliza o método de Runge-Kutta com passo variável (ver Anexo B). O programa foi executado em um computador com processador Intel Core i5-2400 de 3,10GHz, com 8GB de memória RAM DDR3 em 1333MHz, em um sistema operacional Windows 10 de 64bits.

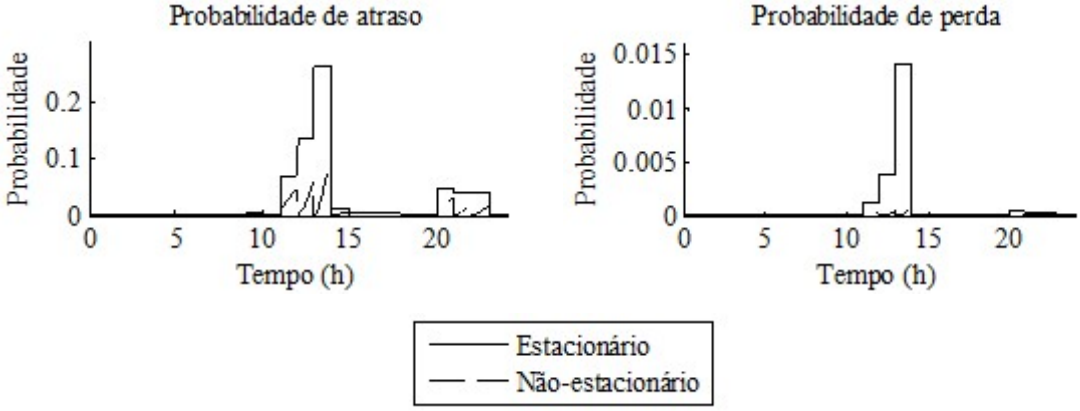
Cada solução para um possível horário de almoço foi calculada, com todas as medidas de desempenho, em cerca de 220 segundos, em média. Para fins de comparação, os resultados obtidos na Seção 5.2, foram calculados em 1,3 segundo, sem jamais ultrapassar 2 segundos. Em média, a solução numérica das equações diferenciais realizou cerca de 1.420 passos (de tamanho variável) ao longo do período. Como o método de Runge-Kutta com passo variável foi escolhido, foi necessário realizar uma interpolação linear em cada solução para encontrar a média dos resultados dos horários das refeições. O processo de interpolação linear e posterior cálculo da média das soluções foi realizado em tempos inferiores à 1 segundo. O autor não afirma que os algoritmos utilizados são os mais eficientes.

Neste cenário original ainda se optou por comparar os resultados entre os modelos estacionário (seguindo a lógica SIPP, explicada na Seção 3.2) e não-estacionário. O objetivo dessa comparação é ilustrar o quão a aproximação estacionária pode distorcer a análise sobre um sistema real.

As primeiras medidas de desempenho obtidas são a probabilidade de perda e o nível de serviço (Equação (52)), observadas na Figura 62. Em ambos casos, a diferença entre os resultados dos modelos é exorbitante, especialmente em pontos críticos. O ponto mais crítico ocorre no almoço, entre às 13 e às 14 horas. O nível de serviço, medido pela probabilidade de atraso, chega à 0,26 no modelo estacionário, enquanto no não-estacionário chega à 0,10, neste período. Quanto à probabilidade de perda, as diferenças relativas são ainda maiores. O modelo

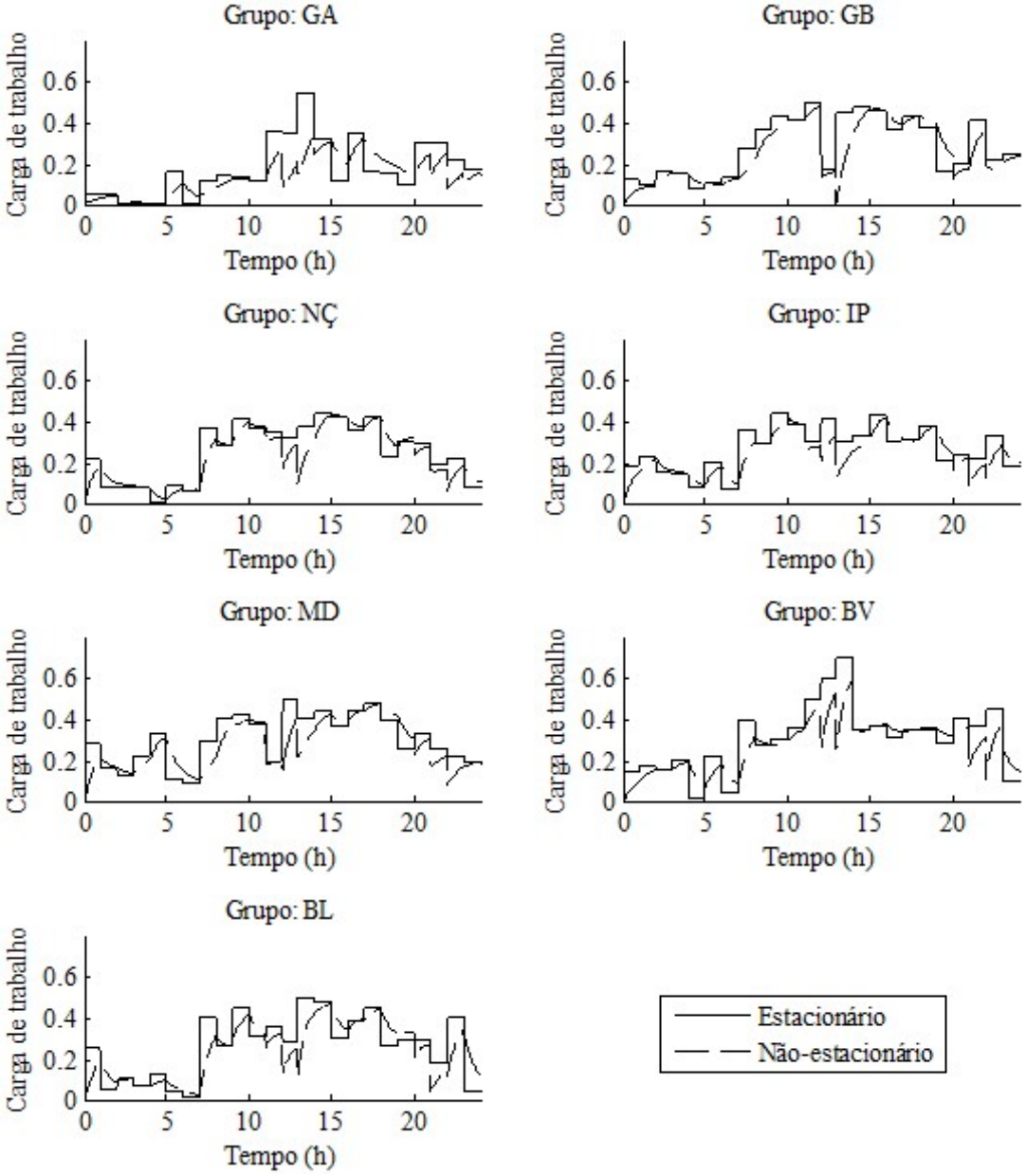
estacionário chega à ordem de 10^{-2}, enquanto o modelo não-estacionário chega à ordem de 10^{-3}.

Figura 62 – Medidas de probabilidade do SAMU-Bauru em seu cenário original.



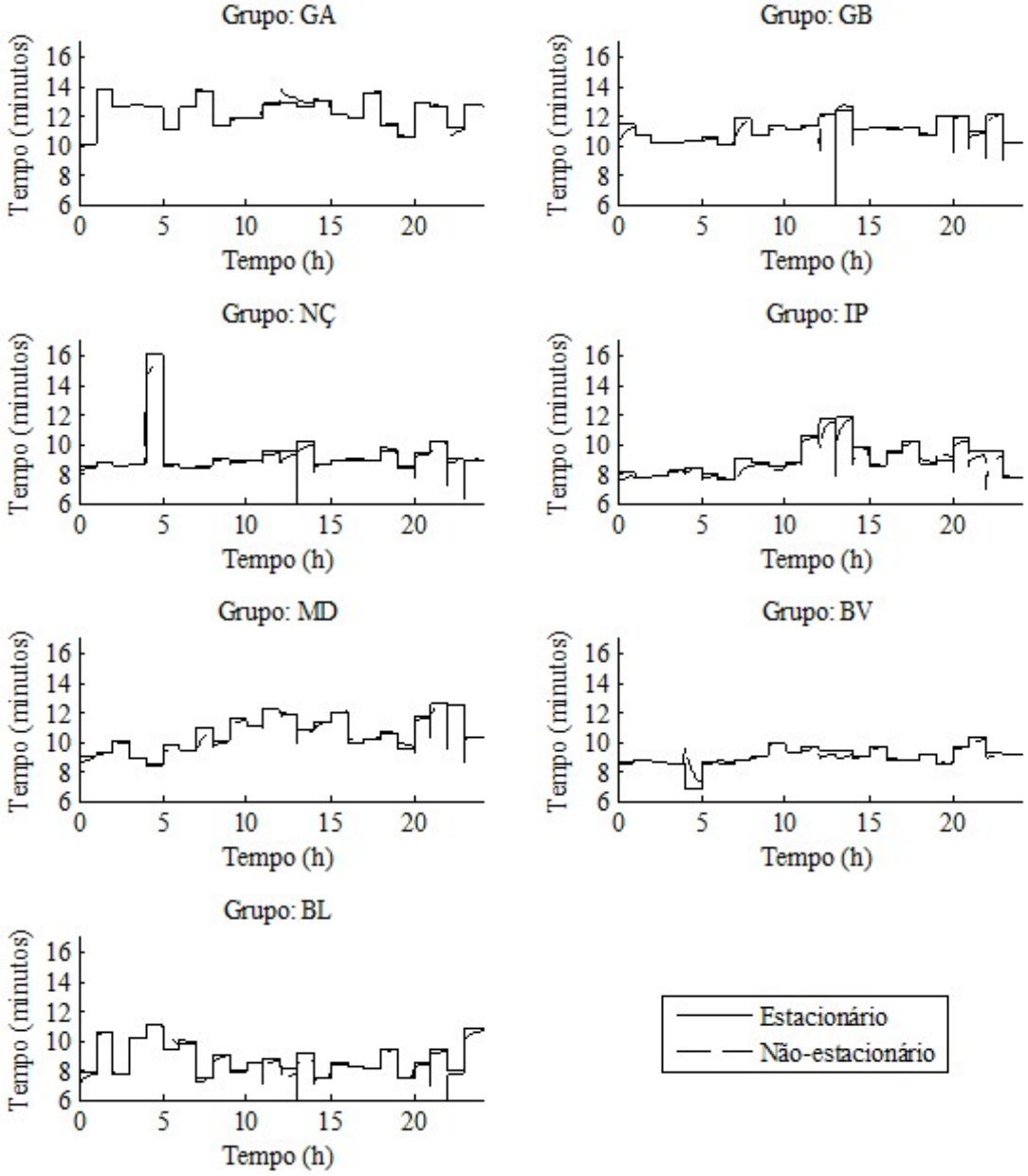
As cargas de trabalho também foram calculadas (Equação (58)) e os resultados para cada grupo de servidores está na Figura 63. Primeiramente, há pontos em que a diferença entre o modelo estacionário e o modelo não-estacionário é pequena. Durante a madrugada, entre 0 e 7 horas, os servidores apresentam uma diferença relativamente pequena em suas cargas de trabalho, o que pode se resultado das baixas taxas de chegada neste período e de não haver nenhuma refeição ou troca de turno neste período. Além disso, o período entre 15 e 19 horas também possui esta mesma característica, neste caso, resultado da pouca variação das taxas de chegada, e da ausência de trocas de servidores no período. Vale ressaltar que esse horário é bastante próximo ao escolhido para a análise estacionária na Seção 5.2. Por outro lado, o horário de almoço se mostra como ponto sensível, especialmente para os servidores dos grupos GA e BV (ambos agrupados). As cargas de trabalho se tornam muito elevadas no período, chegando próximo à 0,6 para o grupo BV. A diferença entre os modelos estacionário e não-estacionário se tornam grandes neste horário para os servidores em geral, com um desvio-relativo médio de 31% no período.

Figura 63 – Cargas de trabalho dos grupos de servidores do SAMU-Bauru em seu cenário original.



Os tempos médios de viagem dos servidores (Equação (63))foram obtidos e são mostrados na Figura 64. É importante ressaltar alguns detalhes sobre o cálculo desta medida de desempenho. Nota-se nos gráficos que os tempos de viagem dos servidores durante os horários de almoço e janta têm distorções instantâneas. Essas distorções são resultadas do processo de interpolação e não prejudicam a análise dos resultados. Os grupos de servidores, como um todo, obtiveram resultados muito próximos entre os modelos estacionário e não-estacionário, mesmo nos horários das refeições. Os desvios-relativos foram menores do que 1% na média, sendo que apenas às 4h houve um desvio superior à 5%.

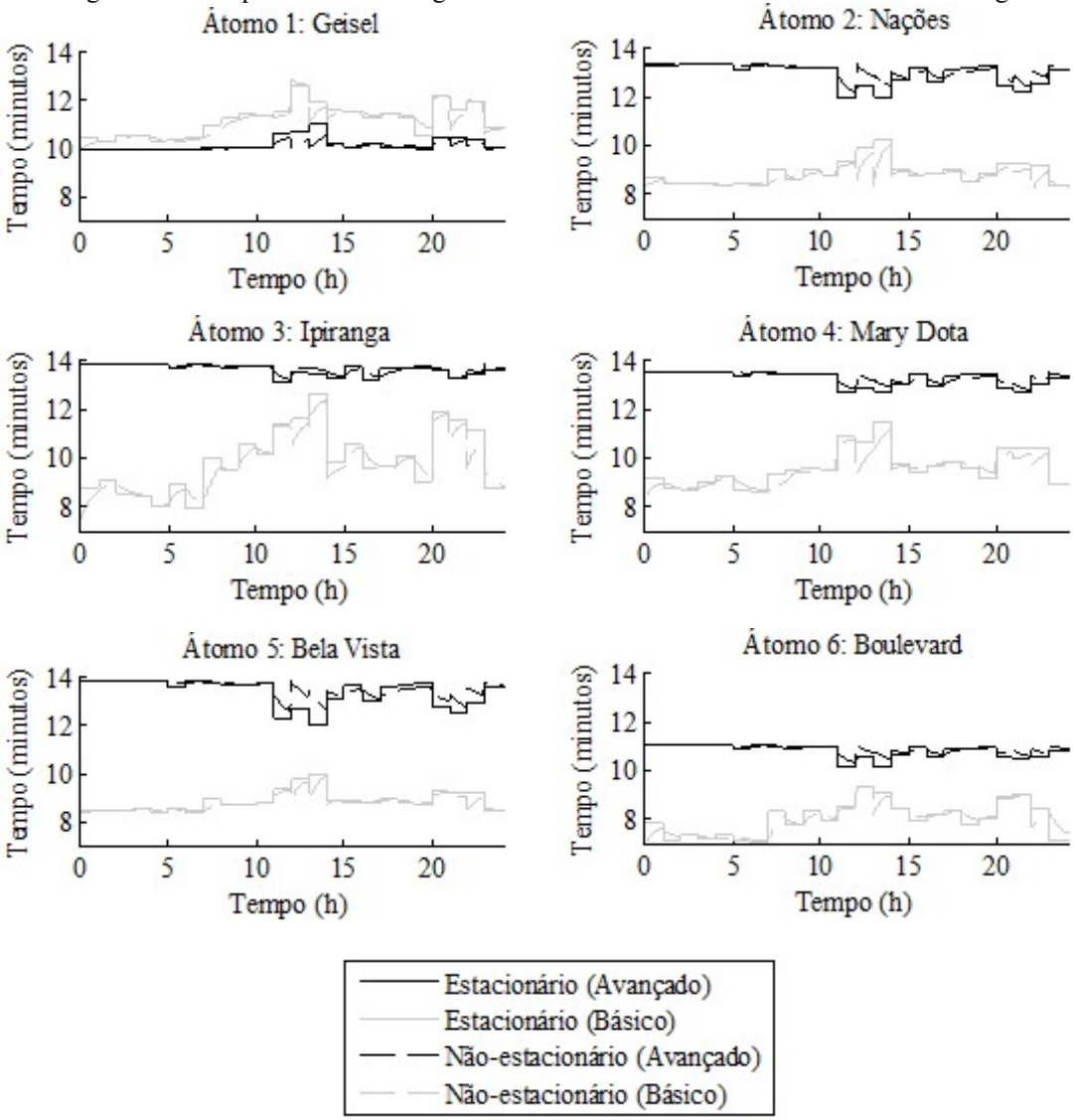
Figura 64 – Tempos médios de viagem dos servidores do SAMU-Bauru em seu cenário original.



A Figura 65 mostra os resultados para os tempos médios de viagem aos átomos (Equação (62)). Com exceção do átomo Geisel, os chamados avançados tiveram sempre um tempo médio de viagem maior do que os chamados básicos. O átomo Geisel é exceção, porque os servidores avançados (grupo GA) estão localizados neste átomo, sendo os chamados básicos atendidos por servidores backups de outros átomos com maior frequência. Os subátomos avançados mostraram uma semelhança grande entre os modelos estacionário e não-estacionário, sendo que o desvio relativo médio foi inferior à 1%, sem jamais ultrapassar 5%. Por outro lado, os subátomos básicos tiveram um desvio relativo médio de 2%, com picos entre 11% e -6%. Em ambos casos, os picos dos desvios são encontrados durante o período de

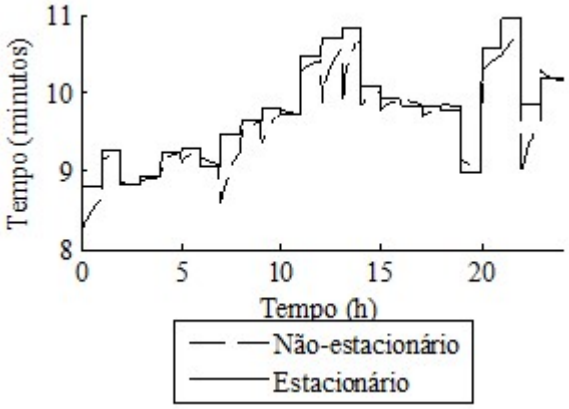
almoço. Esses desvios ainda podem ser considerados pequenos quando comparados aos desvios das cargas de trabalho, do nível de serviço e das probabilidades de perda.

Figura 65 – Tempos médios de viagem aos átomos do SAMU-Bauru em seu cenário original.



A Figura 66 traz o tempo médio de viagem do sistema ao longo do tempo (Equação (61)). O desvio entre os modelos estacionário e não-estacionário possui uma média de 1,5%. Os tempos médios obtidos pelo modelo estacionário são maiores do que os tempos médios obtidos pelo modelo não-estacionário durante praticamente 20 horas do dia. Existem vários picos dos desvios, sendo que cinco deles foram maiores do que 5%, mas nenhum superior à 10%, resultado similar aos obtidos ao se observar os tempos médios de viagem dos servidores e aos átomos.

Figura 66 – Tempo médio de viagem do sistema SAMU-Bauru em seu cenário original.

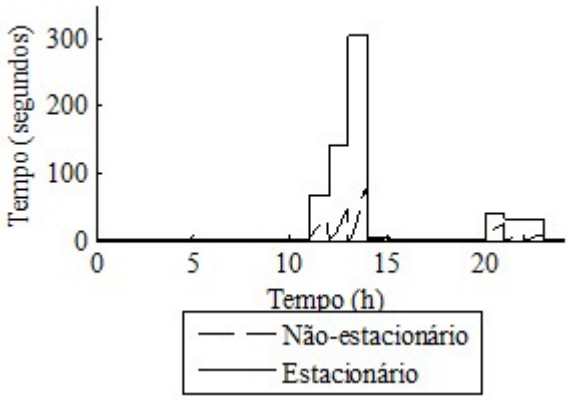


Uma explicação possível para os tempos médios de viagem serem semelhantes em ambos modelos é que os tempos médios de viagem são uma fotografia do instante. Os tempos médios de viagem independem dos eventos futuros, dependem apenas da situação atual do sistema.

Outro ponto a se observar quanto aos tempos médios de viagem é que os tempos médios sofrem saltos e quedas instantâneos, mesmo no modelo não-estacionário em instantes sem troca de turno. Isto pode ser explicado pela mudança sobre as taxas e chegada, alterando os átomos com mais ou menos chamados, alterando também as frequências de usos de *backups*.

A Figura 67 mostra os tempos médios de espera do sistema, calculados através da Equação (31), utilizando a aproximação apresentada na Equação (55). Durante a maior parte do dia, os tempos médios de espera são de poucos segundos, tendo uma diferença pequena entre os modelos estacionário e não-estacionário. Contudo, nos horários das refeições, os tempos médios de espera são mais notáveis. Durante o período de almoço, o modelo estacionário chega a apresentar um tempo médio de espera de aproximadamente 5 minutos, enquanto o modelo não-estacionário apresenta apenas resultado inferior a 1,5 minuto. Durante o período do jantar os tempos médios de espera são menores, chegando a apenas 42 segundos no modelo estacionário e 22 segundos para o modelo não-estacionário.

Figura 67 – Tempo médio de espera do SAMU-Bauru em seu cenário original.



Os resultados apresentados corroboram com as informações coletadas durante a entrevista com os gestores do SAMU-Bauru. Eles se mostraram bastante preocupados com o desempenho do sistema, especialmente, durante o período de almoço das equipes. Seus relatos indicaram que neste período o SAMU-Bauru sofre uma queda na qualidade do serviço oferecido.

Tendo isso em vista, propõe-se o estudo de alguns cenários alternativos. Como forma de reduzir os efeitos do almoço sobre o desempenho do sistema, sugere-se um cenário com horário diferenciado de almoço para as equipes. Outro cenário possível é o aumento na demanda em algum horário específico do dia devido à algum evento não rotineiro.

5.3.3 Resultados para os cenários alternativos

Esta seção fecha a análise do SAMU-Bauru ao longo das 24 horas mostrando os benefícios de se alterar horários das pausas dos servidores e os efeitos que uma onda de calor pode ter sobre o sistema, especialmente durante os momentos mais congestionados.

5.3.3.1 Cenário alternativo 1: alteração horário das refeições

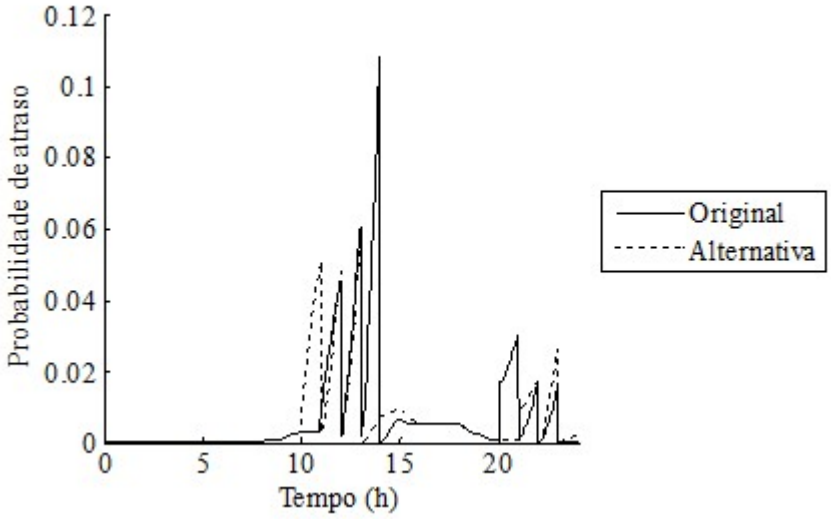
Ao se observar a Figura 61, nota-se que o horário em que são realizadas as refeições das equipes ocorre justamente dentro do período mais congestionado do SAMU-Bauru. O almoço, que se inicia em horários entre às 11 e às 13 horas, é o mais crítico, não por acaso. A última hora de almoço (com início às 13 horas) é a hora de maior demanda do sistema, o que leva aos picos na probabilidade de atraso, no tempo médio de espera observados na Seção 5.3.2. No jantar, que se inicia entre às 20 e às 22h, é possível notar que após às 23 horas (final do horário de jantar) a demanda sobre o SAMU-Bauru cai fortemente.

Por esses motivos, neste cenário, propõe-se alterar o horário das refeições. O horário de almoço foi adiantado em 1 hora e o horário do jantar foi atrasado em 1 hora. O objetivo dessas mudanças é diminuir os efeitos dos horários das refeições sobre o SAMU-Bauru.

As mudanças foram avaliadas dos pontos de vista do gestor e do usuário. Do ponto de vista do gestor, analisou-se o nível de serviço, medido pela probabilidade de atraso. Do ponto de vista do usuário, foram observados o tempo médio de espera e o tempo médio de resposta do sistema.

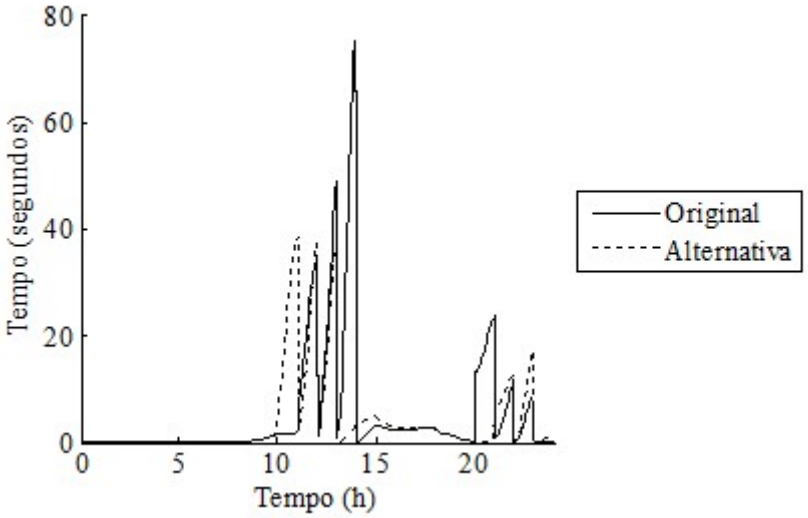
A Figura 68 compara as probabilidades de atraso (Equação (52)) antes e depois da mudança no horário das refeições. A maior diferença observada está na diferença dos picos das probabilidades de atraso. No cenário original o pico da probabilidade de atraso figurava próximo à 0,11, enquanto a alteração do horário de almoço levou o pico para 0,06. Além disso, para o horário do jantar, era possível observar três picos, porém após a mudança do horário do jantar ocorrem apenas dois picos (entre 21 e 23 horas).

Figura 68 – Comparação das probabilidades de atraso entre o cenário original e os horários alternativos das refeições no SAMU-Bauru.



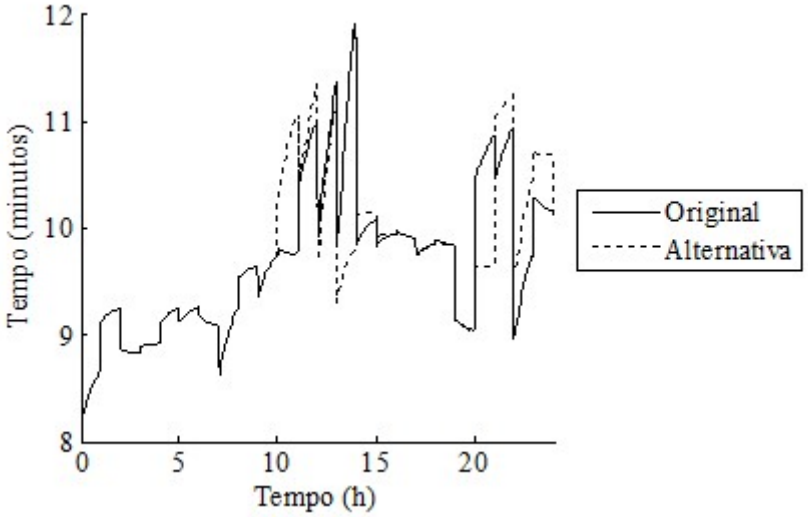
A Figura 69 apresenta a comparação entre os tempos médios de espera do SAMU-Bauru, calculados através da Equação (31), utilizando a aproximação apresentada na Equação (55). Os resultados são bastante parecidos com os obtidos para o nível de serviço. No horário de almoço, o pico chegou próximo a 80 segundos no cenário original. Já no horário alternativo, o pico foi de 40 segundos. No horário do jantar, os picos foram menores, e houve a redução de três para dois picos no horário alternativo com tempos médios de espera de 30 para 20 segundos.

Figura 69 – Comparação dos tempos médios de espera entre o cenário original e os horários alternativos das refeições no SAMU-Bauru.



A Figura 70 compara os tempos médios de resposta do SAMU-Bauru, calculados somando-se os resultados das Equações (61) e (31). A redução dos tempos médios de resposta foi menos sensível do que as reduções vistas nas probabilidades de atraso e nos tempos médios de espera. Durante o horário de almoço, o pico original foi de 12 minutos, contudo a mudanças das refeições levou a uma redução inferior a 1 minuto. O horário do jantar teve um pico próximo ao visto no horário de almoço, próximo aos 11,5 minutos.

Figura 70 – Comparação entre os tempos médios de resposta do SAMU-Bauru no cenário original e com horários alternativos para as refeições.



A alteração das refeições surtiu um efeito benéfico para o sistema, diminuindo os picos de atrasos causados nos horários de maior demanda no SAMU-Bauru. Por outro lado, os tempos

de resposta não mostram um impacto na mesma magnitude. Como ponto mais positivo, a alteração levou o sistema a entrar no horário de pico de demanda (13-14 horas) com todos os servidores disponíveis, obtendo um tempo médio de resposta inferior a 10 minutos.

5.3.3.2 Cenário alternativo 2: aumento na demanda em horário específico

No ano de 2015, o Brasil passou por eventos climáticos bastante adversos, com temperaturas muito elevadas durante o ano. Geirinhas (2016) destaca cidades como Cuiabá e Brasília com temperaturas 9ºC acima da média histórica. Além disso, os jornais têm enfatizado consequências de ondas de calor. Por exemplo, as recentes mortes e queimadas em Portugal (JONES, 2017), a queimada na Espanha que forçou mais de 1.500 pessoas a deixarem suas casas (REUTERS, 2017), e o risco de ondas de calor fatais pelo mundo (MILLER, 2017). Mora *et al.* (2017) mostra o risco fatal das ondas de calor pelo mundo, incluindo alguns dados do Brasil, que em seus modelos é uma região propensa a ondas de calor fatais.

Ondas de calor são caracterizadas pelo aumento na temperatura do ar em uma região por uma sequência razoável de dias (GEIRINHAS, 2016). Elas podem sobrecarregar SAE's visto os problemas de saúde relacionados ao calor. Esses problemas são cãibras (contrações musculares dolorosas), síncope (perda de conhecimento breve), exaustão pelo calor (desidratação, cefaleias, náuseas e vômitos) e golpe de calor (hipertermia e disfunção neurológica central) (MARTO, 2005).

Antes de prosseguir com o cenário, é importante lembrar que ondas de calor são eventos tipicamente do verão. Os dados utilizados para estudo do SAMU-Bauru são do mês de setembro, contudo não houve diferença estatística entre o mês de setembro e os meses de verão (GHUSSN; SOUZA, 2016). Dessa maneira, trabalhar com o impacto das ondas de calor é razoável, vista a época dos dados coletados e as épocas em que o evento é possível de ocorrer. Outro ponto importante a ser citado é que não houve diferença estatística significativa à 5% entre os dias e as semanas dos dados. Isso significa que o sistema não estava sob efeito de ondas de calor durante a coleta dos dados e, portanto, não se está sobrepondo seus efeitos durante a análise.

As ondas de calor ainda possuem uma característica favorável ao estudo utilizando modelos de fila não-estacionários. A característica é a duração do calor por uma sequência de vários dias. Isso é importante na avaliação de um processo de Poisson não-homogêneo (Anexo D), porque os dados não devem variar entre um dia e outro.

Para criar o efeito do aumento na demanda causado pela onda de calor, fez-se com que o aumento fosse maior nos horários mais quentes do dia, entre 14 e 16 horas. Além disso, seus efeitos já podem ser sentidos a partir das 10 horas, horário usual em que se recomenda evitar o sol e o calor intenso no verão do Brasil.

Para representar o aumento crescente na demanda entre 10 e 15 horas e depois o aumento decrescente entre 15 e 20 horas, utilizou-se proporções de uma pirâmide de Pascal em sua linha 9. A Equação (64) mostra os elementos da referida linha da pirâmide de Pascal.

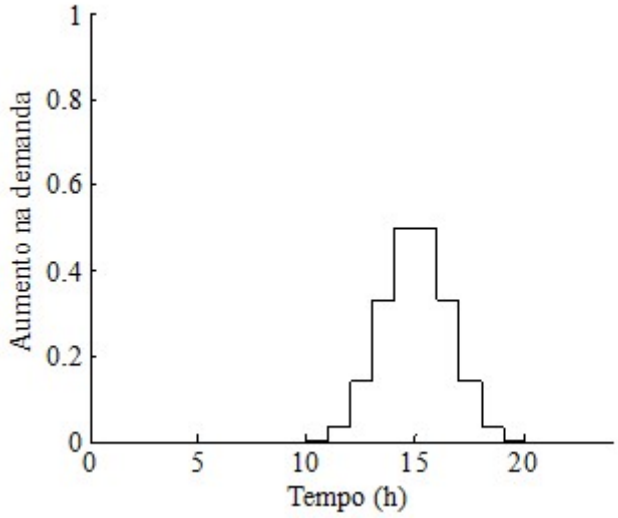
$$\text{Linha } 9 \quad [1 \quad 9 \quad 36 \quad 84 \quad 126 \quad 126 \quad 84 \quad 36 \quad 9 \quad 1] \tag{64}$$

O aumento máximo na demanda foi escolhido, arbitrariamente, como 50%. Assim, os pontos máximos do aumento na demanda são aqueles com os maiores valores da linha da Pirâmide de Pascal, os elementos 126. Os demais elementos têm um aumento proporcional a este, como mostra a Equação (65).

$$0{,}5\left[\frac{1}{126} \quad \frac{9}{126} \quad \frac{36}{126} \quad \frac{84}{126} \quad \frac{126}{126} \quad \frac{126}{126} \quad \frac{84}{126} \quad \frac{36}{126} \quad \frac{9}{126} \quad \frac{1}{126}\right] \tag{65}$$

Cada elemento da Equação (65) representa o aumento na demanda em um horário do dia. De forma que, entre 10 e 11 horas, o aumento na demanda é de $\frac{1}{126}0{,}5$; entre 11 e 12 horas, é de $\frac{9}{126}0{,}5$; e assim por diante até as 20 horas, como mostra a Figura 71.

Figura 71 – Simulação do aumento na demanda do SAMU-Bauru causado por uma onda de calor.

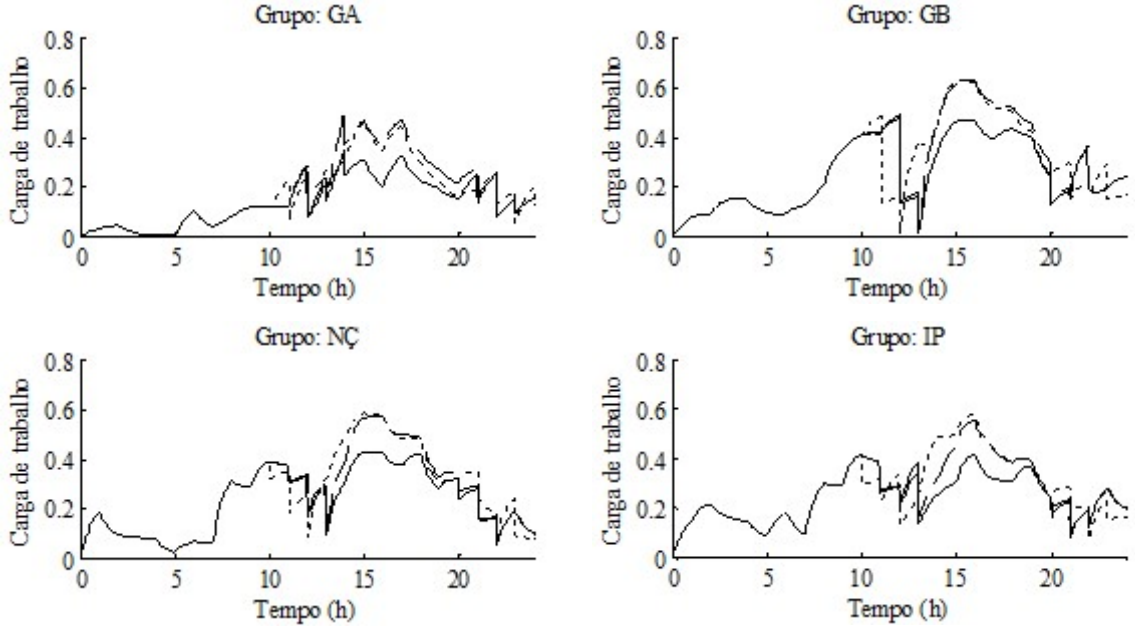


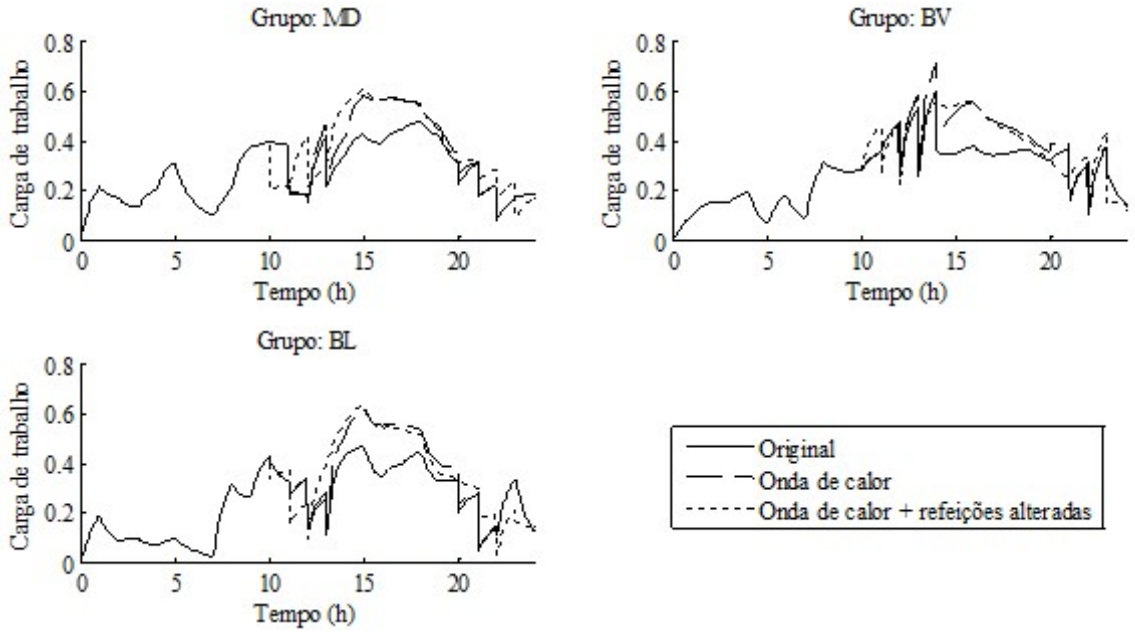
Para compreender os efeitos de uma onda de calor sobre o SAMU-Bauru, foram observadas medidas de desempenho internas e externas. A medida interna, do ponto de vista do gestor, escolhida foi a carga de trabalho. Já a medida externa, do ponto de vista do usuário, escolhida foi o tempo médio de espera. As medidas de desempenho foram comparadas entre

três cenários: o cenário original; sob o efeito de uma onda de calor; e sob o efeito de uma onda de calor com o horário das refeições alteradas.

A Figura 72 compara as cargas de trabalho dos servidores (Equação (58)). Primeiramente, é possível observar que a maioria dos servidores chegam a ficar 60% do seu tempo ocupados durante o horário mais severo da onda de calor, aproximadamente. Isso significa, que os servidores ficam 45% mais tempo ocupados em relação ao cenário original. As exceções são os VSA's, do grupo GA, que ficam 50% do tempo ocupados, uma carga inferior à média observada para os VSB's. A ambulâncias mais sensíveis ao horário de almoço foram os grupos GA e BV. Adiantar o almoço ajuda a evitar o pico do aumento na demanda do horário mais quente. Além disso, adiantar o almoço reduziu a carga de trabalho até o final do dia em todos servidores.

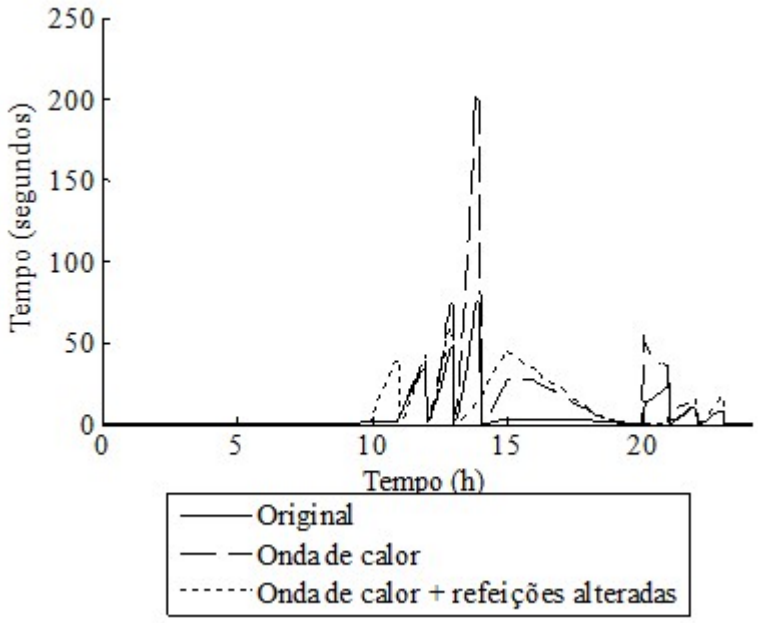
Figura 72 – Comparação das cargas de trabalho do SAMU-Bauru para o cenário de uma onda de calor.





A Figura 73 compara os tempos médios de espera do SAMU-Bauru, calculados utilizando a Equação (31), utilizando a aproximação apresentada na Equação (55). Quando apenas sob o efeito da onda de calor, o horário de almoço se tornou ainda mais crítico com um pico no tempo médio de espera de 205 segundos, ao invés dos 75 segundos do cenário original. Por outro lado, adiantar o almoço em uma hora diminui o pico da espera para quase 60 segundos. No horário mais quente do dia, 15 horas, há um pico local de tempos de espera de 28 segundos, mas quando se adianta o almoço este pico local é de 45 segundos. Isso pode ser explicado pela volta do horário de almoço dos servidores. Apenas sob o efeito da onda de calor, há servidores tornando disponíveis em horários mais próximos das 15 horas, diminuindo este pico local. Quando se adianta o almoço, estes servidores já estão trabalhando a mais tempo. Não é visto como vantagem para a operação ter o pico local reduzido às custas de um pico global muito mais alto.

Figura 73 – Comparação dos tempos médios de espera do SAMU-Bauru para o cenário de uma onda de calor.



Estes resultados mostram que adiantar o almoço em uma hora ainda é vantajoso. Os picos globais do tempo médio de espera são atenuados, assim como a carga de trabalho é reduzida até o final do dia, após o aumento na demanda causado pela onda de calor.

5.4 O SAMU-Bauru em 2017

Conforme mencionado ao longo da Seção 5.2.3, o átomo Nações foi a melhor opção para a inclusão de uma nova ambulância. Contudo, entre o ano de 2013 (ano dos dados coletados) e o ano de 2017, o SAMU-Bauru passou por algumas alterações em sua configuração e a base localizada no átomo Nações deixou de ser operada, assim como a base Boulevard. Os VSB's destas bases foram transferidos para os átomos Geisel e Ipiranga, respectivamente. Estas bases, embora possuam localizações privilegiadas para o atendimento enfrentavam problemas de infraestrutura. Estas bases eram vulneráveis à enchentes e alagamentos tornando-as, ou inoperantes em vários períodos, ou com operação parcial durante os períodos de chuva intensa.

Outra mudança que ocorreu neste período, foi a aquisição de novas ambulâncias para o funcionamento de uma frota dupla. O número de equipes trabalhando não se alterou, sendo 2 VSA's e 7 VSB's, mas as novas ambulâncias foram adquiridas para facilitar o rodízio das ambulâncias na realização das manutenções, já que quebras eram constantes nas ambulâncias, prejudicando o atendimento. Além disso, a frota dupla auxilia durante as trocas de turno, já que

as equipes iniciando o turno não precisam esperar o retorno das ambulâncias em atividade para começarem a operar.

Ainda não foi possível utilizar o modelo hipercubo não-estacionário para avaliar o quanto estas mudanças foram positivas para o SAMU-Bauru. O modelo ainda não possui calibragem dos tempos de serviço, ferramenta essencial para se trabalhar com cenários em que há alterações nos servidores. Também por isso, não foram estudados cenários com a inclusão ou exclusão de servidores pelos mesmos motivos.

6 CONCLUSÕES

Este trabalho propôs um estudo em duas etapas de um SAE utilizando o modelo hipercubo de filas. A primeira etapa focou na melhoria em nível tático do SAMU, observando o impacto do aumento na demanda e encontrando a melhor localização para um servidor novo nessas condições. A segunda etapa focou em uma melhoria em nível operacional, analisando possível mudança nos horários para as refeições das equipes e suas vantagens.

Para preparar o modelo para a primeira etapa, foram consideradas características já estudadas na literatura como prioridade em fila e aleatoriedade no despacho. Além dessas características o SAMU-Bauru opera com suas ambulâncias avançadas reservadas para o atendimento de chamados graves. A reserva dessas ambulâncias foi chamada de reserva total de capacidade para se diferenciar da reserva de capacidade vista em Iannoni *et al.* (2015), onde os servidores não foram diferenciados e a reserva levou em conta apenas as classes dos chamados. O modelo hipercubo foi estendido para lidar com o impedimento completo desses servidores atenderem aos chamados básicos. A fila foi dividida de acordo com as prioridades e a cauda pode ser formada mesmo com os servidores avançados livres. Ainda se aproveitando da aleatoriedade no despacho, da homogeneidade de servidores e da co-localização dos mesmos, foi possível obter ganhos computacionais sem perda na precisão dos resultados. Isso foi possível estendendo o modelo hipercubo considerando grupos de servidores, chamada de agregação de servidores, alterando o espaço de estados do modelo hipercubo, para não diferenciar servidores desnecessariamente.

Para preparar o modelo para a segunda etapa, foram consideradas as alterações que o SAMU-Bauru sofre ao longo do dia de operação. O SAMU-Bauru opera com equipes de dois turnos de 12 horas, cujo início se dá às 7 e às 19 horas. As chegadas de chamados também variam, atingindo seu pico durante a tarde e seu mínimo antes do amanhecer. As refeições das equipes também afetam a operação do sistema, retirando, momentaneamente, equipes de suas operações. Os horários das refeições são escolhidos e informados à central dentro de um intervalo de tempo de 2 horas (11-13 horas e 20-22 horas). A variação temporal do SAMU-Bauru foi modelada considerando um modelo estacionário independente período a período (SIPP) e considerando um modelo não-estacionário baseado nas equações de Chapman-Kolmogorov. O modelo não-estacionário foi expandido para considerar disciplinas de fim de turno exaustiva (sem interrupção do atendimento) e preemptiva (com interrupção do atendimento). As refeições foram incluídas no modelo seguindo as regras de pausa do SAMU-Bauru, gerando 10 cenários com pausas fixas para as refeições.

Os dados foram coletados do SAMU-Bauru referentes ao mês de setembro de 2013. A primeira etapa da análise considerou apenas os chamados do período de pico, identificado entre 12 e 18 horas (GHUSSN; SOUZA, 2016). A segunda etapa considerou todos os chamados ao longo do dia, considerados aproximadamente constantes de hora em hora.

A análise em duas etapas foi importante por possibilitar a realização de análises tanto do comportamento médio do sistema, como das variações horárias que o sistema enfrenta diariamente. A reserva de capacidade não introduziu novas medidas de desempenho, mas uma maneira de analisar uma política de despacho diferenciada. Beojone e Souza (2017) realizou estudo semelhante, porém, por não se utilizar da reserva total de capacidade, seus resultados foram tiveram uma aderência inferior aos aqui obtidos. Por outro lado, a análise não-estacionária abriu caminho para o estudo de quaisquer variações horárias sofridas por SAE's. Por exemplo, as pausas para as refeições representaram os períodos de maior instabilidade nas medidas de desempenho do SAMU-Bauru, confirmando os relatos dos gestores do sistema.

A análise também previu o estudo de cenários alternativos do SAMU-Bauru. Durante a primeira etapa, foram discutidos aumentos na demanda do SAMU-Bauru e a melhor localização para a inclusão de uma nova ambulância. O aumento na demanda foi obtido por meio do estudo de séries temporais. Durante a segunda etapa, foram abordadas situações com a alteração do horário das refeições e os efeitos de um aumento na demanda durante o horário mais congestionado do dia. A causa do aumento na demanda no horário mais congestionado, período da tarde, seria uma onda de calor, como a que afetou a o continente europeu no ano de 2017.

Algumas dificuldades e limitações foram enfrentadas ao longo do processo de pesquisa. Primeiramente, as hipóteses não tiveram total aderência à amostra, especialmente no tocante aos tempos de serviço e às taxas de chegada no modelo não-estacionário, ou precisaram ser relaxadas. Os desvios médios encontrados foram, em geral, inferiores à 10% durante a primeira etapa. O tamanho da amostra se mostrou um desafio a parte quando feita a análise não-estacionária. A amostra inicial com mais de 600 registros foi separada hora a hora, assim, apenas 12 dos 24 intervalos observados tiveram uma amostra com mais de 30 registros. A agregação dos dados de maneira a diminuir o número de intervalos estudados não foi utilizada, visto que isto eliminaria as variações percebidas entre certos horários do dia. O tamanho da amostra se mostrou um problema para verificar diferenças nos tempos de serviço das ambulâncias, tornando necessário agregar os dados de um turno completo para a amostra ser de um tamanho razoável. Como os dados dos chamados são inseridos manualmente em um formulário em computador, alguns erros e inconsistências foram encontrados. Foram menos do que 5% dos chamados da amostra com essas ocorrências, mas que poderiam ter sido evitados

caso o sistema já trabalhasse automaticamente com o uso de GPS. Souza (2010) já considerou o processo de implementação dessas ferramentas nos SAMU's, porém após mais de 5 anos isso ainda não foi observado no SAMU-Bauru. Quarto, por ainda não possuir calibragem dos tempos de serviço, não foi possível estudar cenários com alterações na configuração dos servidores. Por exemplo, não foi possível comparar a configuração atual do SAMU-Bauru com a de 2013. Também não é possível trabalhar com mudanças buscando melhorias, como a inclusão de uma ambulância durante o turno mais congestionado, ou a retirada durante o turno menos congestionado. Quinto, a representação dos horários de refeições foi feita através da geração de cenários com pausas em horários fixos, contudo as pausas são realizadas probabilisticamente, com duração fixa em 1 hora. Apesar desse método possibilitar a comparação com o modelo estacionário, a análise das pausas se torna menos confiável e precisa. É importante ressaltar que este problema ocorre apenas para a disciplina de fim de turno exaustiva, já que os servidores podem continuar trabalhando após o horário da pausa. Isso se torna um problema diferente do fim de turno, visto que a pausa é muito curta e é altamente provável que o servidor não tenha seu tempo de pausa respeitado. Não foram encontrados métodos para este tipo de representação na literatura de teoria das filas, mas pode ser encontrado em softwares de simulação de eventos discretos (INGOLFSSON, 2005).

Algumas perspectivas interessantes para pesquisas futuras podem ser levantadas a partir das dificuldades e limitações encontradas.

Uma primeira perspectiva seria a realização de um estudo de caso a fim de encontrar barreiras, desafios e melhores práticas para o processo de implantação de métodos automáticos de coleta de dados das ambulâncias nos SAMU's pelo Brasil. Isso seria útil para alavancar a expansão dessa ferramenta pelo Brasil, melhorando a qualidade dos dados obtidos dos SAMU's brasileiros e, assim, melhorar as análises realizadas neles.

Outra perspectiva é o desenvolvimento de estudos utilizando a análise em duas etapas, mas especialmente a análise não estacionária em SAE's mais congestionados. Buscando dessa maneira uma amostra de dados maior para a realização de todos os testes necessários, assim como a melhoria das ferramentas para a realização desses testes.

Para melhorar o modelo hipercubo não-estacionário, propõe-se o estudo de expandi-lo de forma que trabalhe com calibragem dos tempos de serviço. A calibragem poderia ser feita a cada instante, considerando os resultados do instante anterior.

Uma linha de pesquisa alternativa seria o estudo e desenvolvimento de métodos para representar pausas curtas em sistemas de disciplina exaustiva, como o horário de almoço no SAMU-Bauru. O desafio é representar situações em que o servidor para após o término do

atendimento atual por um tempo prefixado (1 hora por exemplo) em um modelo de teoria das filas.

Outra linha de pesquisa relacionada às paradas dos servidores, seria o estudo de uma representação formal de pausas sem horário fixo, que ocorrem dentro de um intervalo de tempo. O desafio é trabalhar para que o vetor de probabilidade do sistema sofra as alterações gradualmente ao longo do tempo em que as pausas podem ser iniciadas.

Também seria interessante estudar o uso dos modelos de cadeia de Markov discreta para representar a alteração da localização de uma ambulância, adicionando-a a um outro grupo já existente no sistema. Tal cadeia de Markov poderia ser representada por uma matriz bastante semelhante à utilizada para a transição de estado instantânea da disciplina de fim de turno preemptiva. Esse modelo pode funcionar de maneira semelhante aos modelos que utilizam uma tabela de conformidade, como Alanis *et al.* (2013) e van Barneveld *et al.* (2017).

Como não é garantido que os algoritmos utilizados sejam os mais eficientes, é sugerida a busca por métodos e algoritmos mais eficientes. A ideia é que se tire proveito de características do sistema para melhorar seu tempo computacional. Além disso, sugere-se a busca por outros métodos dentro da pesquisa operacional para se trabalhar com os problemas estudados, ou outros semelhantes.

REFERÊNCIAS BIBLIOGRÁFICAS

ALANIS, R.; INGOLFSSON, A.; KOLFAL, B. A Markov Chain Model for an EMS System with Repositioning. **Production and Operations Management**, 22, n. 1, 2013. 216-231.

ARENALES, M. et al. **Pesquisa Operacional:** para cursos de engenharia. 2ª. ed. Rio de Janeiro: Elsevier: ABEPRO, 2015.

ATKINSON, J. B. et al. Heuristic methods for the analysis of a queuing system describing emergency medical service deployed along a highway. **Cybernetics and Systems Analysis**, 42, n. 3, 2006. 379-391.

BASSAMBOO, A.; HARRISON, J. M.; ZEEVI, A. Design and Control of a Large Call Center: Assymptotic Analysis of an LP-Based Method. **Operations Research**, 54, n. 3, 2006. 419-435.

BATTA, R.; DOLAN, J.; KRISHNAMURTHY, N. The Maximal Expected Covering Location Problem: Revisited. **Transportation Science**, 23, n. 4, 1989. 277-287.

BEOJONE, C. V.; SOUZA, R. M. Application of the hypercube model with queue priorities and more than one preferential server: a case study on a SAMU. **Gestão & Produção**, São Carlos, In Press, 2017.

BERTSIMAS, D.; MOURTZINOU, G. Transient laws of non-stationary queueing systems and their applications. **Queueing Systems**, 25, 1997. 115-155.

BOYACI, B.; GEROLIMINIS, N. Approximation methods for large-scale spatial queueing systems. **Transportation Research Part B**, 74, 2015. 151-181.

BOYCE, W. E.; DIPRIMA, R. C. **Equações Diferenciais Elementares e Problemas de Valores de Contorno**. Tradução de V M Iorio. 7ª. ed. Rio de Janeiro: LTC Editora, 2001. 416p p.

BRAILSFORD, S.; VISSERS, J. OR in healthcare: A European perspective. **European Journal of Operational Research**, 212, 2011. 223-234.

BRANDEAU, M.; LARSON, R. Extending and applying the hypercube queueing model to deploy ambulances in Boston. In: SWERSEY, A. J.; INGNALL, E. J. **Delivery of Urban Services:** TIMS Studies in the Management Science 22. [S.l.]: Elsevier, 1986. p. 121-153.

BROWN, L. et al. Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective. **Journal of the American Statistical Association**, 100, n. 469, 2005. 36-50.

BUFFA, E. S.; COSGROVE, M. J.; LUCE, B. J. An integrated work shift scheduling system. **Decision Sciences**, 7, n. 4, 1976. 620-630.

BURWELL, T.; JARVIS, J.; MCKNEW, M. Modeling co-located servers and dispatch ties in hypercube model. **Computers & Operations Research**, 20, n. 2, 1993. 113-119.

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

CHIYOSHI, F. Y.; GALVÃO, R. D.; MORABITO, R. A note on solutions to the maximal expected covering location problem. **Computers & Operations Research**, 30, 2003. 87-96.

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

CHIYOSHI, F.; GALVÃO, R.; MORABITO, R. Modelo hipercubo: análise e resultados para o caso de servidores não-homogêneos. **Pesquisa Operacional**, 21, n. 2, 2001. 199-218.

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

CHUNG, K.; MIN, D. Staffing a service system with appointment based customer arrival. **Journal of the Operational Research Society**, 65, 2014. 1533-1543.

CONSELHO FEDERAL DE MEDICINA (CFM). Resolução CFM Nº 2.149/2016. In: BRASIL **Diário Oficial da União**. Brasília: Seção I, 2016. p. 99.

DANTZIG, G. B. A Comment on Edie's "Traffic Delays at Toll Booths". **Journal of the Operations Research Society of America**, 2, n. 3, 1954. 339-341.

DASKIN, M. S. A Maximum Expected Covering Location Model: Formulation, Properties and Heuristic Solution. **Transportation Science**, 17, n. 1, 1983. 48-70.

DAVOUDPOUR, H.; MORTAZ, E.; HOSSEINIJOU, S. A new probabilistic coverage model for ambulances deployment with hypercube queuing approach. **International Journal of Advanced Manufacturing Technology**, 70, 2014. 1157-1168.

DEFRAEYE, M.; NIEUWENHUYSE, I. V. Staffing and Scheduling under nonstationary demand for service: A literature review. **Omega**, 58, 2016. 4-25.

DIETZ, D. Practical scheduling for call center operations. **Omega**, 39, 2011. 550-557.

EDIE, L. C. Traffic Delays at Toll Booths. **Journal of the Operations Research Society of America**, 2, n. 2, 1954. 107-138.

EDWARDS, C. H.; PENNEY, D. E. **Equações Diferenciais Elementares:** com Problemas de Contorno. Tradução de C Wilmer; L B Castro e P Viana. 3ª. ed. Rio de Janeiro: Prentice-Hall, 1995. 643p p.

EHLERS, R. S. **Análise de Séries Temporais**. 1ª. ed. [S.l.]: Departamento de Estatística, UFPR, 2005.

ERKUT, E.; INGOLFSSON, A.; ERDOGAN, G. Ambulance Location for Maximum Survival. **Naval Research Logistics**, 55, 2007. 42-58.

FELDMAN, Z. et al. Staffing of Time-Varying Queues to Achieve Time-Stable Performance. **Management Science**, 54, n. 2, 2008. 324-338.

GALVÃO, R. D. et al. Solução do problema de localização de máxima disponibilidade utilizando o modelo hipercubo. **Pesquisa Operacional**, 23, n. 1, 2003. 61-78.

GALVÃO, R. D.; CHIYOSHI, F. Y.; MORABITO, R. Towards unified formulations and extensions of two classical probabilistic location models. **Computers & Operations Research**, 32, 2005. 15-33.

GALVÃO, R. D.; MORABITO, R. Emergency service systems: the use of the hypercube queueing model in the solution of probabilistic location problems. **International Transactions in Operational Research**, 15, 2008. 525-549.

GANS, N.; KOOLE, G.; MANDELBAUM, A. Telephone Call Centers: Tutorial, Review and Research Prospects. **Manufacturing & Service Operations Management**, 5, n. 2, 2003. 79-141.

GEIRINHAS, J. L. M. Caracterização Climática e Sinóptica das Ondas de Calor no Brasil. **Dissertação de Mestrado**, Universidade de Lisboa, Lisboa, Portugal, p. 73, 2016.

GEROLIMINIS, N.; KARLAFTIS, M.; SKABARDONIS, A. A spatial queuing model for the emergency vehicle districting and location problem. **Transportation Research Part B**, 43, 2009. 798-811.

GEROLIMINIS, N.; KEPAPTSOGLOU, K.; KARLAFTIS, M. A hybrid hypercube - Genetic algorithm approach for deploying many emergency response mobile units in an urban network. **European Journal of Operational Research**, 210, 2011. 287-300.

GHUSSN, L.; SOUZA, R. M. Análise de desempenho do SAMU/Bauru-SP em períodos de pico de demanda. **Gestão da Produção, Operações e Sistemas**, Bauru, 11, n. 3, 2016. 75-103.

GILLARD, J.; KNIGHT, V. Using Singular Spectrum Analysis to obtain staffing level requirements in emergency units. **Journal of the Operational Research Society**, 65, 2014. 735-746.

GRASSMANN, W. Transient solutions in Markovian queues: an algorithm for finding them and determining their waiting-time distributions. **European Journal of Operational Research**, 1, n. 6, 1977. 396-402.

GREEN, L. V.; KOLESAR, P. J.; SOARES, J. An improved heuristic for staffing telephone call centers with limited operating hours. **Production and Operations Management**, 12, n. 1, 2003. 46-61.

GREEN, L. V.; KOLESAR, P. J.; WHITT, W. Coping with Time-Varying Demand When Setting Staffing Requirements for a Service System. **Production and Operations Management**, 16, n. 1, 2007. 13-39.

GREEN, L.; KOLESAR, P.; SOARES, J. Improving the Sipp Approach for Staffing Service Systems That Have Cyclic Demands. **Operations Research**, 49, n. 4, 2001. 549-564.

GREEN, L.; SOARES, J. Note-Computing time-dependent waiting time probabilities in M(t)/M/s(t) Queueing systems. **Manufacturing & Service Operations Management**, 9, n. 1, 2007. 54-61.

GURVICH, I.; ARMORY, M.; MANDELBAUM, A. Service-Level Differentiation in Call Centers with Fully Flexible Servers. **Management Science**, 54, n. 2, 2008. 297-294.

HULSHOF, P. et al. Analytical models to determine room requirements in outpatient clinics. **OR Spectrum**, 34, 2012. 391-405.

IANNONI, A. P. Otimização da configuração e operação de sistemas médico emergenciais em rodovias utilizando o modelo hipercubo. **Tese de Doutorado**, Universidade Federal de São Carlos, São Carlos, 2005.

IANNONI, A. P.; CHIYOSHI, F.; MORABITO, R. A spatially distributed queuing model considering dispatching policies with server reservation. **Transportation Research Part E**, 75, 2015. 46-66.

IANNONI, A. P.; MORABITO, R. Modelo de fila hipercubo com múltiplo despacho e backup parcial para análise de sistemas de atendimento emergencial em rodovias. **Pesquisa Operacional**, 26, n. 3, 2006. 493-519.

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

IANNONI, A.; MORABITO, R. A multiple dispatch and partial backup hypercube queueing model to analyze emergency medical systems on highways. **Transportation Research Part E**, 43, 2008. 755-771.

IANNONI, A.; MORABITO, R.; SAYDAM, C. An optimization approach for ambulance location and the districting of the response segments on highways. **European Journal of Operational Research**, 195, 2009. 528-542.

INGOLFSSON, A. Modeling the M(t)/M/s(t) Queue with Exhaustive Discipline, Edmonton, 2005. Disponivel em: <http://www.bus.ualberta.ca/aingolfsson/working_papers.htm>.

INGOLFSSON, A. et al. A Suvey and Experimental Comparison of Service-Level-Approximation Methods for Nonstationary M(t)/M/s(t) Queueing Systems with Exhaustive Discipline. **INFORMS Journal on Computing**, 19, n. 2, 2007. 201-214.

INGOLFSSON, A. et al. Combining integer programming and the randomization method to schedule employees. **European Journal of Operational Research**, 202, 2010. 153-163.

INGOLFSSON, A.; BUDGE, S.; ERKUT, E. Optimal ambulance location with random delays and travel times. **Health Care Management Science**, 11, 2008. 262-274.

INGOLFSSON, A.; HAQUE, A.; UMNIKOV, A. Accounting for time-varying queueing effects in workforce scheduling. **European Journal of Operational Research**, 139, 2002. 585-597.

JARVIS, J. Approximating the equilibrium behavior of multi-server loss systems. **Management Science**, 31, n. 2, 1985. 235-239.

JCNET. Em Bauru, Samu regional vai integrar 16 cidades. **Revista Emergência**, Abril 2010. Disponivel em: <http://www.revistaemergencia.com.br/site/content/noticias/noticia_detalhe.php?id=AJy4Jy>. Acesso em: 14 Maio 2016.

JENNINGS, O. B. et al. Server Staffing to meet time-varying demanda. **Management Science**, 42, n. 10, 1996. 1383-1394.

JONES, S. Huge forest fires in Portugal kill at least 60. **The Guardian: World**, 18 Junho 2017. Disponivel em: <https://www.theguardian.com/world/2017/jun/18/portugal-more-than-20-people-killed-in-forest-fires>. Acesso em: 5 jul. 2017.

JUNG, M.; LEE, E. S. Numerical Optimization of a Queueing System by Dynamic Programming. **Journal of Mathematical Analysis and Applications**, 141, 1989. 84-93.

KEITH, E. G. Operator Scheduling. **AIIE Transactions**, 11, n. 1, 1979. 37-41.

KENDALL, D. G. Stochastic Processes Occuring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain. **The Annals of Mathematical Statistics**, 24, n. 3, 1953. 338-354.

KIM, D. S.; SMITH, R. L. An Exact Aggregation/Disaggregation Algorithm for Large Scale Markov Chains. **Naval Research Logistics**, 42, n. 7, 1995. 1115-1128.

KIM, S.-H.; WHITT, W. Are Call Centers and Hospital Arrivals Well Modeled by Nonhomogeneous Poisson Processes. **Manufacturing & Service Operations Management**, 16, n. 3, 2014. 464-480.

KIMBER, R. M.; MARLOW, M.; HOLLIS, E. M. Flow/Delay relantionships for major/minor priority junctions. **Traffic Engineering and Control**, 18, n. 11, 1977. 516-519.

KLEINROCK, L. **Queueing Systems:** Theory. [S.l.]: Wiley, v. I, 1975.

KOLESAR, P. J. et al. A Queuing-Linear Programming Approach to Scheduling Police Patrol Cars. **Operations Research**, 23, n. 6, 1975. 1045-1062.

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

LARSON, R. Approximating the Performance of Urban Emergency Service Systems. **Operations Research**, 23, n. 5, 1975. 845-868.

LARSON, R.; ODONI, A. **Urban Operations Research**. [S.l.]: Prentice-Hall, 2007. Disponivel em: <http://web.mit.edu/urban_or_book/www/book/>.

LEWIS, P. A. W. Some Results on Tests for Poisson Processes. **Biometrika**, 52, n. 1/2, 1965. 67-77.

LIU, N.; D'AUNNO, T. The Productivity and Cost-Efficiency of Models for Involving Nurse Practioners in Primary Care: A Perspective from Queueing Analysis. **Health Services Research**, 47, n. 2, 2012. 594-613.

LOPES, S. L. B.; FERNANDES, R. J. Uma breve revisão do atendimento médico pré-hospitalar. **Simpósio: Trauma II**, Medicina, Ribeirão Preto, 1999. 381-387.

LUQUE, L. Um estudo de métodos de solução do modelo hipercubo de filas para sistemas de grande porte. **Dissertação de Mestrado**, INPE, São José dos Campos, 2008. 147.

MANDELBAUM, A.; ZELTYN, S. Staffing Many-Server Queues with Impatient Customers: Constraint Satisfaction in Call Centers. **Operations Research**, 57, n. 5, 2009. 1189-1205.

MARGOLIUS, B. H. Transient solution to the time-dependent multiserver Poisson queue. **Journal of Applied Probability**, 42, n. 3, 2005. 766-777.

MARIANOV, V.; REVELLE, C. The Queueing Maximal Avaiability Location Problem: A model for the siting of emergency vehicles. **European Journal of Operational Research**, 93, 1996. 110-120.

MARTO, N. Ondas de Calor: Impacto sobre a saúde. **Acta Médica Potuguesa**, 18, 2005. 467-474.

MASSEY, W. A.; WHITT, W. Peak congestion in multi-server service systems with slow varying arrival rates. **Queueing Systems**, 25, 1997. 157-172.

MAXWELL, M. S. et al. Approximate Dynamic Programming for Ambulance Redeployment. **INFORMS Journal on Computing**, 22, n. 2, 2010. 266-281.

MENDONÇA, F.; MORABITO, R. Analysing emergency medical service ambulance deployment on a Brazilian using the hypercube queueing model. **Journal of the Operational Research Society**, 52, 2001. 261-270.

MILLER, B. Deadly heat waves becoming more common due to climate change. **CNN: World**, 20 Junho 2017. Disponivel em: <http://edition.cnn.com/2017/06/19/world/killer-heat-waves-rising/index.html>. Acesso em: 5 jul. 2017.

MINITAB INC. **Getting Started with Minitab 17**. [S.l.]: [s.n.], 2010. Disponivel em: <www.minitab.com>.

MORA, C. et al. Global risk of deadly heat. **Nature Climate Change**, 7, n. 7, 2017. 1-7.

MORABITO, R.; CHIYOSHI, F.; GALVÃO, R. Non-homogeneous servers in emergency medical systems: Practical applications using the hypercube queueing model. **Socio-Economic Planning Sciences**, 42, 2008. 255-270.

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

PATRICK, J.; PUTERMAN, M. L.; QUEYRANNE, M. Dynamic Multipriority Patient Scheduling for a Diagnostic Resource. **Operations Research**, 56, n. 6, 2008. 1507-1525.

PRONOVOST, P. J. et al. Framework for Patient Safety Research and Improvement. **Circulation**, 119, n. 2, 2009. 330-337.

RAJAGOPALAN, H. K.; SAYDAM, C.; XIAO, J. A multiperiod set covering location model for dynamic redeployment of ambulances. **Computers & Operations Research**, 35, 2008. 814-826.

REUTERS. Spain fire forces more than 1,500 from homes and campsites. **The Guardian: World**, 25 Junho 2017. Disponivel em:

<https://www.theguardian.com/world/2017/jun/25/spain-forest-fire-forces-more-than-1500-from-homes-and-campsites>. Acesso em: 5 jul. 2017.

REVELLE, C. S.; EISELT, H. A. Location analysis: A syntesis and survey. **European Journal of Operational Research**, 165, 2005. 1-19.

REVELLE, C.; HOGAN, K. The Maximum Availability Location Problem. **Transportation Science**, 23, n. 3, 1989. 192-200.

RODRIGUES, L. F. et al. Towards hypercube queueing models for dispatch policies with priority in queue and partial backup. **Computers & Operations Research**, 84, 2017. 92-105.

ROTHKOPF, M. H.; OREN, S. S. A closure approximation for the nonstationary M/M/s queue. **Management Science**, 25, n. 6, 1979. 522-534.

SACKS, S. R.; LARSON, R. C.; SCHAACK, C. Minimizing the Cost of Dispatch Delays by Holding Patrol Cars in Reserve. **Journal of Quantitative Criminology**, 9, n. 2, 1993. 203-224.

SAYDAM, C.; AYTUG, H. Accurate estimation of expected coverage: revisited. **Socio-Economic Planning Sciences**, 37, 2003. 69-80.

SAYDAM, C.; REPEDE, J.; BURWELL, T. Accurate Estimation of Expected Coverage: A Comparative Study. **Socio-Economic Planning Sciences**, 28, n. 2, 1994. 113-120.

SCHMID, V. Solving the dynamic ambulance relocation and dispatching problemusing approximate dynamic programming. **European Journal of Operational Research**, 219, 2012. 611-621.

SCHWARZ, J.; SELINKA, G.; STOLLETZ, R. Performance analysis of time-dependent queueing systems: Survey and classification. **Omega**, 63, 2016. 170-189.

SIMPSON, N.; HANCOCK, P. Fifty years of operational research and emergency response. **Journal of the Operational Research Society**, 60, 2009. S126-S139.

SOUZA, R. et al. Incorporating priorities for waiting customers in the hypercube queuing model with application to an emergency medical service system in Brazil. **European Journal of Operational Research**, 242, 2015. 274-285.

SOUZA, R. M. Análise da configuração de SAMU utilizando modelo hipercubo com prioridade na fila e múltiplas alternativas de localização de ambulâncias. **Tese de doutorado**, UFSCAR, São Carlos, 2010.

SOUZA, R. M. et al. Análise da configuração de SAMU utilizando múltiplas alternativas de localização de ambulâncias. **Gestão & Produção**, 20, n. 2, 2013. 287-302.

STOLLETZ, R. Approximation of the non-stationary M(t)/M(t)/c(t)-queue using stationary queueing models: the stationary backlog-carryover approach. **European Journal of Operational Research**, 190, 2008. 478-493.

STOLLETZ, R.; LAGERSHAUSEN, S. Time-dependent performance evaluation for loss-waiting queues with arbitrary distributions. **International Journal of Production Research**, 51, n. 5, 2013. 1366-1378.

TAHA, H. A. **Pesquisa Operacional:** uma visão geral. Tradução de A S Marques. 8ª. ed. São Paulo: Prentice-Hall, 2008. 346p p.

TAKEDA, R. A. Uma contribuição para avaliar o desempenho de sistemas de transporte emergencial de saúde. **Tese de doutorado**, USP, São Carlos, 2000.

TAKEDA, R.; WIDMER, J.; MORABITO, R. Aplicação do modelo hipercubo de filas para avaliar a descentralização de ambulâncias em um sistema urbano de atendimento médio de urgência. **Pesquisa Operacional**, 24, n. 1, 2004. 39-72.

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

TAYLOR, I. D. S.; TEMPLETON, J. G. C. Waiting Time in a Multi-Server Cutoff-Priority Queue, and Its Application on an Urban Ambulance Service. **Operations Research**, 28, n. 5, 1980. 1168-1188.

THOMPSON, G. M. Labor Staffing and Scheduling Models for Controlling Service Levels. **Naval Research Logistics**, 44, 1997. 719-740.

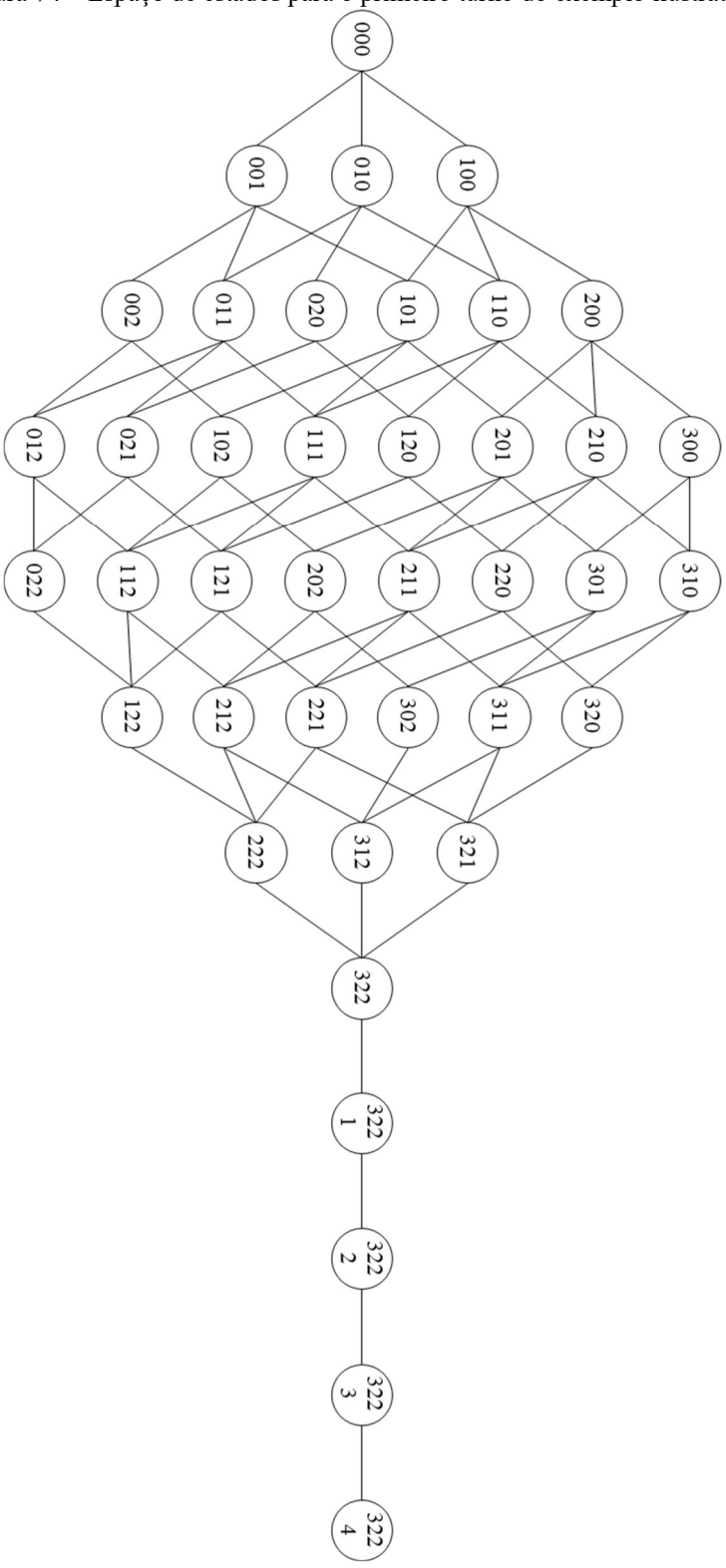
VAN BARNEVELD, T. C.; VAN DER MEI, R. D.; BHULAI, S. Compliance tables for an EMS system with two types of medical response units. **Computers and Operations Research**, 80, 2017. 68-81.

WHITT, W. What you should know about queueing models to set staffing requirements in service systems. **Naval Research Logistics**, 54, n. 5, 2007. 476-484.

APÊNDICE A

A Figura 74 traz o espaço de estados e transições para o primeiro turno do exemplo ilustrativo da Seção 4.3.

Figura 74 – Espaço de estados para o primeiro turno do exemplo ilustrativo.



As transições de estado causadas a partir da chegada de um novo chamado ocorrem como segue:

- $\{000\} \rightarrow \{100\}$: quando ocorre uma chegada do átomo 1, com taxa λ_1;
- $\{000\} \rightarrow \{010\}$: quando ocorre uma chegada do átomo 2, com taxa λ_2;
- $\{000\} \rightarrow \{001\}$: quando ocorre uma chegada do átomo 3, com taxa λ_3;
- $\{100\} \rightarrow \{200\}$: quando ocorre uma chegada do átomo 1, com taxa λ_1;
- $\{100\} \rightarrow \{110\}$: quando ocorre uma chegada do átomo 2, com taxa λ_2;
- $\{100\} \rightarrow \{101\}$: quando ocorre uma chegada do átomo 3, com taxa λ_3;
- $\{010\} \rightarrow \{110\}$: quando ocorre uma chegada do átomo 1, com taxa λ_1;
- $\{010\} \rightarrow \{020\}$: quando ocorre uma chegada do átomo 2, com taxa λ_2;
- $\{010\} \rightarrow \{011\}$: quando ocorre uma chegada do átomo 3, com taxa λ_3;
- $\{001\} \rightarrow \{101\}$: quando ocorre uma chegada do átomo 1, com taxa λ_1;
- $\{001\} \rightarrow \{011\}$: quando ocorre uma chegada do átomo 2, com taxa λ_2;
- $\{001\} \rightarrow \{002\}$: quando ocorre uma chegada do átomo 3, com taxa λ_3;
- $\{200\} \rightarrow \{300\}$: quando ocorre uma chegada do átomo 1, com taxa λ_1;
- $\{200\} \rightarrow \{210\}$: quando ocorre uma chegada do átomo 2, com taxa λ_2;
- $\{200\} \rightarrow \{201\}$: quando ocorre uma chegada do átomo 3, com taxa λ_3;
- $\{110\} \rightarrow \{210\}$: quando ocorre uma chegada do átomo 1, com taxa λ_1;
- $\{110\} \rightarrow \{120\}$: quando ocorre uma chegada do átomo 2, com taxa λ_2;
- $\{110\} \rightarrow \{111\}$: quando ocorre uma chegada do átomo 3, com taxa λ_3;
- $\{101\} \rightarrow \{201\}$: quando ocorre uma chegada do átomo 1, com taxa λ_1;
- $\{101\} \rightarrow \{111\}$: quando ocorre uma chegada do átomo 2, com taxa λ_2;
- $\{101\} \rightarrow \{102\}$: quando ocorre uma chegada do átomo 3, com taxa λ_3;
- $\{020\} \rightarrow \{120\}$: quando ocorre uma chegada do átomo 1, com taxa λ_1;
- $\{020\} \rightarrow \{021\}$: quando ocorre uma chegada dos átomos 2 ou 3, com taxa $\lambda_2 + \lambda_3$;
- $\{011\} \rightarrow \{111\}$: quando ocorre uma chegada do átomo 1, com taxa λ_1;
- $\{011\} \rightarrow \{021\}$: quando ocorre uma chegada do átomo 2, com taxa λ_2;
- $\{011\} \rightarrow \{012\}$: quando ocorre uma chegada do átomo 3, com taxa λ_3;
- $\{002\} \rightarrow \{102\}$: quando ocorre uma chegada do átomo 1, com taxa λ_1;
- $\{002\} \rightarrow \{012\}$: quando ocorre uma chegada dos átomos 2 ou 3, com taxa $\lambda_2 + \lambda_3$;
- $\{300\} \rightarrow \{310\}$: quando ocorre uma chegada do átomo 2 ou metade chegadas do átomo 1, com taxa $\lambda_1/2 + \lambda_2$;

- $\{300\} \rightarrow \{301\}$: quando ocorre uma chegada do átomo 3 ou metade chegadas do átomo 1, com taxa $\lambda_1/2 + \lambda_3$;
- $\{210\} \rightarrow \{310\}$: quando ocorre uma chegada do átomo 1, com taxa λ_1;
- $\{210\} \rightarrow \{220\}$: quando ocorre uma chegada do átomo 2, com taxa λ_2;
- $\{210\} \rightarrow \{211\}$: quando ocorre uma chegada do átomo 3, com taxa λ_3;
- $\{201\} \rightarrow \{301\}$: quando ocorre uma chegada do átomo 1, com taxa λ_1;
- $\{201\} \rightarrow \{211\}$: quando ocorre uma chegada do átomo 2, com taxa λ_2;
- $\{201\} \rightarrow \{202\}$: quando ocorre uma chegada do átomo 3, com taxa λ_3;
- $\{120\} \rightarrow \{220\}$: quando ocorre uma chegada do átomo 1, com taxa λ_1;
- $\{120\} \rightarrow \{121\}$: quando ocorre uma chegada dos átomos 2 ou 3, com taxa $\lambda_2 + \lambda_3$;
- $\{111\} \rightarrow \{211\}$: quando ocorre uma chegada do átomo 1, com taxa λ_1;
- $\{111\} \rightarrow \{121\}$: quando ocorre uma chegada do átomo 2, com taxa λ_2;
- $\{111\} \rightarrow \{112\}$: quando ocorre uma chegada do átomo 3, com taxa λ_3;
- $\{102\} \rightarrow \{202\}$: quando ocorre uma chegada do átomo 1, com taxa λ_1;
- $\{102\} \rightarrow \{112\}$: quando ocorre uma chegada dos átomos 2 ou 3, com taxa $\lambda_2 + \lambda_3$;
- $\{021\} \rightarrow \{121\}$: quando ocorre uma chegada do átomo 1, com taxa λ_1;
- $\{021\} \rightarrow \{022\}$: quando ocorre uma chegada dos átomos 2 ou 3, com taxa $\lambda_2 + \lambda_3$;
- $\{012\} \rightarrow \{112\}$: quando ocorre uma chegada do átomo 1, com taxa λ_1;
- $\{012\} \rightarrow \{022\}$: quando ocorre uma chegada dos átomos 2 ou 3, com taxa $\lambda_2 + \lambda_3$;
- $\{310\} \rightarrow \{320\}$: quando ocorre uma chegada do átomo 2 ou metade chegadas do átomo 1, com taxa $\lambda_1/2 + \lambda_2$;
- $\{310\} \rightarrow \{311\}$: quando ocorre uma chegada do átomo 3 ou metade chegadas do átomo 1, com taxa $\lambda_1/2 + \lambda_3$;
- $\{301\} \rightarrow \{311\}$: quando ocorre uma chegada do átomo 2 ou metade chegadas do átomo 1, com taxa $\lambda_1/2 + \lambda_2$;
- $\{301\} \rightarrow \{302\}$: quando ocorre uma chegada do átomo 3 ou metade chegadas do átomo 1, com taxa $\lambda_1/2 + \lambda_3$;
- $\{220\} \rightarrow \{320\}$: quando ocorre uma chegada do átomo 1, com taxa λ_1;
- $\{220\} \rightarrow \{221\}$: quando ocorre uma chegada dos átomos 2 ou 3, com taxa $\lambda_2 + \lambda_3$;
- $\{211\} \rightarrow \{311\}$: quando ocorre uma chegada do átomo 1, com taxa λ_1;
- $\{211\} \rightarrow \{221\}$: quando ocorre uma chegada do átomo 2, com taxa λ_2;
- $\{211\} \rightarrow \{212\}$: quando ocorre uma chegada do átomo 3, com taxa λ_3;

- $\{202\} \rightarrow \{302\}$: quando ocorre uma chegada do átomo 1, com taxa λ_1;
- $\{202\} \rightarrow \{212\}$: quando ocorre uma chegada dos átomos 2 ou 3, com taxa $\lambda_2 + \lambda_3$;
- $\{121\} \rightarrow \{221\}$: quando ocorre uma chegada do átomo 1, com taxa λ_1;
- $\{121\} \rightarrow \{122\}$: quando ocorre uma chegada dos átomos 2 ou 3, com taxa $\lambda_2 + \lambda_3$;
- $\{112\} \rightarrow \{212\}$: quando ocorre uma chegada do átomo 1, com taxa λ_1;
- $\{112\} \rightarrow \{122\}$: quando ocorre uma chegada dos átomos 2 ou 3, com taxa $\lambda_2 + \lambda_3$;
- $\{022\} \rightarrow \{122\}$: quando ocorre uma chegada de qualquer átomo, com taxa λ;
- $\{320\} \rightarrow \{321\}$: quando ocorre uma chegada de qualquer átomo, com taxa λ;
- $\{311\} \rightarrow \{321\}$: quando ocorre uma chegada do átomo 2 ou metade chegadas do átomo 1, com taxa $\lambda_1/2 + \lambda_2$;
- $\{311\} \rightarrow \{312\}$: quando ocorre uma chegada do átomo 3 ou metade chegadas do átomo 1, com taxa $\lambda_1/2 + \lambda_3$;
- $\{302\} \rightarrow \{312\}$: quando ocorre uma chegada de qualquer átomo, com taxa λ;
- $\{221\} \rightarrow \{321\}$: quando ocorre uma chegada do átomo 1, com taxa λ_1;
- $\{221\} \rightarrow \{222\}$: quando ocorre uma chegada dos átomos 2 ou 3, com taxa $\lambda_2 + \lambda_3$;
- $\{212\} \rightarrow \{312\}$: quando ocorre uma chegada do átomo 1, com taxa λ_1;
- $\{212\} \rightarrow \{222\}$: quando ocorre uma chegada dos átomos 2 ou 3, com taxa $\lambda_2 + \lambda_3$;
- $\{122\} \rightarrow \{222\}$: quando ocorre uma chegada de qualquer átomo, com taxa λ;
- $\{321\} \rightarrow \{322\}$: quando ocorre uma chegada de qualquer átomo, com taxa λ;
- $\{312\} \rightarrow \{322\}$: quando ocorre uma chegada de qualquer átomo, com taxa λ;
- $\{222\} \rightarrow \{322\}$: quando ocorre uma chegada de qualquer átomo, com taxa λ;
- $\{322\} \rightarrow \{322|1\}$: quando ocorre uma chegada de qualquer átomo, com taxa λ;
- $\{322|1\} \rightarrow \{322|2\}$: quando ocorre uma chegada de qualquer átomo, com taxa λ;
- $\{322|2\} \rightarrow \{322|3\}$: quando ocorre uma chegada de qualquer átomo, com taxa λ; e
- $\{322|3\} \rightarrow \{322|4\}$: quando ocorre uma chegada de qualquer átomo, com taxa λ;

As transições de estado causadas a partir da finalização de um serviço ocorrem como segue:

- $\{100\} \rightarrow \{000\}$: quando um servidor do grupo 1 termina um serviço com taxa μ_1;
- $\{010\} \rightarrow \{000\}$: quando um servidor do grupo 2 termina um serviço com taxa μ_2;
- $\{001\} \rightarrow \{000\}$: quando um servidor do grupo 3 termina um serviço com taxa μ_3;
- $\{200\} \rightarrow \{100\}$: quando um servidor do grupo 1 termina um serviço com taxa $2\mu_1$;

- $\{110\} \rightarrow \{100\}$: quando um servidor do grupo 2 termina um serviço com taxa μ_2;
- $\{110\} \rightarrow \{010\}$: quando um servidor do grupo 1 termina um serviço com taxa μ_1;
- $\{101\} \rightarrow \{100\}$: quando um servidor do grupo 3 termina um serviço com taxa μ_3;
- $\{101\} \rightarrow \{001\}$: quando um servidor do grupo 1 termina um serviço com taxa μ_1;
- $\{020\} \rightarrow \{010\}$: quando um servidor do grupo 2 termina um serviço com taxa $2\mu_2$;
- $\{011\} \rightarrow \{010\}$: quando um servidor do grupo 3 termina um serviço com taxa μ_3;
- $\{011\} \rightarrow \{001\}$: quando um servidor do grupo 2 termina um serviço com taxa μ_2;
- $\{002\} \rightarrow \{001\}$: quando um servidor do grupo 3 termina um serviço com taxa $2\mu_3$;
- $\{300\} \rightarrow \{200\}$: quando um servidor do grupo 1 termina um serviço com taxa $3\mu_1$;
- $\{210\} \rightarrow \{200\}$: quando um servidor do grupo 2 termina um serviço com taxa μ_2;
- $\{210\} \rightarrow \{110\}$: quando um servidor do grupo 1 termina um serviço com taxa $2\mu_1$;
- $\{201\} \rightarrow \{200\}$: quando um servidor do grupo 3 termina um serviço com taxa μ_3;
- $\{201\} \rightarrow \{101\}$: quando um servidor do grupo 1 termina um serviço com taxa $2\mu_1$;
- $\{120\} \rightarrow \{110\}$: quando um servidor do grupo 2 termina um serviço com taxa $2\mu_2$;
- $\{120\} \rightarrow \{020\}$: quando um servidor do grupo 1 termina um serviço com taxa μ_1;
- $\{111\} \rightarrow \{110\}$: quando um servidor do grupo 3 termina um serviço com taxa μ_3;
- $\{111\} \rightarrow \{101\}$: quando um servidor do grupo 2 termina um serviço com taxa μ_2;
- $\{111\} \rightarrow \{011\}$: quando um servidor do grupo 1 termina um serviço com taxa μ_1;
- $\{021\} \rightarrow \{020\}$: quando um servidor do grupo 3 termina um serviço com taxa μ_3;
- $\{021\} \rightarrow \{011\}$: quando um servidor do grupo 2 termina um serviço com taxa $2\mu_2$;
- $\{012\} \rightarrow \{011\}$: quando um servidor do grupo 3 termina um serviço com taxa $2\mu_3$;
- $\{012\} \rightarrow \{002\}$: quando um servidor do grupo 2 termina um serviço com taxa μ_2;
- $\{310\} \rightarrow \{300\}$: quando um servidor do grupo 2 termina um serviço com taxa μ_2;
- $\{310\} \rightarrow \{210\}$: quando um servidor do grupo 1 termina um serviço com taxa $3\mu_1$;
- $\{301\} \rightarrow \{300\}$: quando um servidor do grupo 3 termina um serviço com taxa μ_3;
- $\{301\} \rightarrow \{201\}$: quando um servidor do grupo 1 termina um serviço com taxa $3\mu_1$;
- $\{220\} \rightarrow \{210\}$: quando um servidor do grupo 2 termina um serviço com taxa $2\mu_2$;
- $\{220\} \rightarrow \{120\}$: quando um servidor do grupo 1 termina um serviço com taxa $2\mu_1$;
- $\{211\} \rightarrow \{210\}$: quando um servidor do grupo 3 termina um serviço com taxa μ_3;
- $\{211\} \rightarrow \{201\}$: quando um servidor do grupo 2 termina um serviço com taxa μ_2;
- $\{211\} \rightarrow \{111\}$: quando um servidor do grupo 1 termina um serviço com taxa $2\mu_1$;
- $\{202\} \rightarrow \{201\}$: quando um servidor do grupo 3 termina um serviço com taxa $2\mu_3$;

- $\{202\} \rightarrow \{102\}$: quando um servidor do grupo 1 termina um serviço com taxa $2\mu_1$;
- $\{121\} \rightarrow \{120\}$: quando um servidor do grupo 3 termina um serviço com taxa μ_3;
- $\{121\} \rightarrow \{111\}$: quando um servidor do grupo 2 termina um serviço com taxa $2\mu_2$;
- $\{121\} \rightarrow \{021\}$: quando um servidor do grupo 1 termina um serviço com taxa μ_1;
- $\{112\} \rightarrow \{111\}$: quando um servidor do grupo 3 termina um serviço com taxa $2\mu_3$;
- $\{112\} \rightarrow \{102\}$: quando um servidor do grupo 2 termina um serviço com taxa μ_2;
- $\{112\} \rightarrow \{012\}$: quando um servidor do grupo 1 termina um serviço com taxa μ_1;
- $\{022\} \rightarrow \{021\}$: quando um servidor do grupo 3 termina um serviço com taxa $2\mu_3$;
- $\{022\} \rightarrow \{012\}$: quando um servidor do grupo 2 termina um serviço com taxa $2\mu_2$;
- $\{320\} \rightarrow \{310\}$: quando um servidor do grupo 2 termina um serviço com taxa $2\mu_2$;
- $\{320\} \rightarrow \{220\}$: quando um servidor do grupo 1 termina um serviço com taxa $3\mu_1$;
- $\{311\} \rightarrow \{310\}$: quando um servidor do grupo 3 termina um serviço com taxa μ_3;
- $\{311\} \rightarrow \{301\}$: quando um servidor do grupo 2 termina um serviço com taxa μ_2;
- $\{311\} \rightarrow \{211\}$: quando um servidor do grupo 1 termina um serviço com taxa $3\mu_1$;
- $\{302\} \rightarrow \{301\}$: quando um servidor do grupo 3 termina um serviço com taxa $2\mu_3$;
- $\{302\} \rightarrow \{202\}$: quando um servidor do grupo 1 termina um serviço com taxa $3\mu_1$;
- $\{221\} \rightarrow \{220\}$: quando um servidor do grupo 3 termina um serviço com taxa μ_3;
- $\{221\} \rightarrow \{211\}$: quando um servidor do grupo 2 termina um serviço com taxa $2\mu_2$;
- $\{221\} \rightarrow \{121\}$: quando um servidor do grupo 1 termina um serviço com taxa $2\mu_1$;
- $\{212\} \rightarrow \{211\}$: quando um servidor do grupo 3 termina um serviço com taxa $2\mu_3$;
- $\{212\} \rightarrow \{202\}$: quando um servidor do grupo 2 termina um serviço com taxa μ_2;
- $\{212\} \rightarrow \{112\}$: quando um servidor do grupo 1 termina um serviço com taxa $2\mu_1$;
- $\{122\} \rightarrow \{121\}$: quando um servidor do grupo 3 termina um serviço com taxa $2\mu_3$;
- $\{122\} \rightarrow \{112\}$: quando um servidor do grupo 2 termina um serviço com taxa $2\mu_2$;
- $\{122\} \rightarrow \{022\}$: quando um servidor do grupo 1 termina um serviço com taxa μ_1;
- $\{321\} \rightarrow \{320\}$: quando um servidor do grupo 3 termina um serviço com taxa μ_3;
- $\{321\} \rightarrow \{311\}$: quando um servidor do grupo 2 termina um serviço com taxa $2\mu_2$;
- $\{321\} \rightarrow \{221\}$: quando um servidor do grupo 1 termina um serviço com taxa $3\mu_1$;
- $\{312\} \rightarrow \{311\}$: quando um servidor do grupo 3 termina um serviço com taxa $2\mu_3$;
- $\{312\} \rightarrow \{302\}$: quando um servidor do grupo 2 termina um serviço com taxa μ_2;
- $\{312\} \rightarrow \{212\}$: quando um servidor do grupo 1 termina um serviço com taxa $3\mu_1$;
- $\{222\} \rightarrow \{221\}$: quando um servidor do grupo 3 termina um serviço com taxa $2\mu_3$;

- $\{222\} \rightarrow \{212\}$: quando um servidor do grupo 2 termina um serviço com taxa $2\mu_2$;
- $\{222\} \rightarrow \{122\}$: quando um servidor do grupo 1 termina um serviço com taxa $2\mu_1$;
- $\{322\} \rightarrow \{321\}$: quando um servidor do grupo 3 termina um serviço com taxa $2\mu_3$;
- $\{322\} \rightarrow \{312\}$: quando um servidor do grupo 2 termina um serviço com taxa $2\mu_2$;
- $\{322\} \rightarrow \{222\}$: quando um servidor do grupo 1 termina um serviço com taxa $3\mu_1$;
- $\{322|1\} \rightarrow \{322\}$: quando qualquer servidor termina um serviço com uma taxa μ;
- $\{322|2\} \rightarrow \{322|1\}$: quando qualquer servidor termina um serviço com uma taxa μ;
- $\{322|3\} \rightarrow \{322|2\}$: quando qualquer servidor termina um serviço com uma taxa μ;
- $\{322|4\} \rightarrow \{322|3\}$: quando qualquer servidor termina um serviço com uma taxa μ;

As equações diferenciais que representam todas as transições de chegada e de serviço deste sistema estão representadas pelas Equações (66)-(103).

$$P'_{\{000\}}(t) = -\lambda P_{\{000\}}(t) + \mu_1 P_{\{100\}}(t) + \mu_2 P_{\{010\}}(t) + \mu_3 P_{\{001\}}(t) \quad (66)$$

$$P'_{\{100\}}(t) = -(\lambda + \mu_1) P_{\{100\}}(t) + \lambda_1 P_{\{000\}}(t) + 2\mu_1 P_{\{200\}}(t) + \mu_2 P_{\{110\}}(t) + \mu_3 P_{\{101\}}(t) \quad (67)$$

$$P'_{\{010\}}(t) = -(\lambda + \mu_2) P_{\{010\}}(t) + \lambda_2 P_{\{000\}}(t) + \mu_1 P_{\{110\}}(t) + 2\mu_2 P_{\{020\}}(t) + \mu_3 P_{\{011\}}(t) \quad (68)$$

$$P'_{\{001\}}(t) = -(\lambda + \mu_3) P_{\{001\}}(t) + \lambda_3 P_{\{000\}}(t) + \mu_1 P_{\{101\}}(t) + \mu_2 P_{\{011\}}(t) + 2\mu_3 P_{\{002\}}(t) \quad (69)$$

$$P'_{\{200\}}(t) = -(\lambda + 2\mu_1) P_{\{200\}}(t) + \lambda_1 P_{\{100\}}(t) + 3\mu_1 P_{\{300\}}(t) + \mu_2 P_{\{210\}}(t) + \mu_3 P_{\{201\}}(t) \quad (70)$$

$$P'_{\{110\}}(t) = -(\lambda + \mu_1 + \mu_2) P_{\{110\}}(t) + \lambda_1 P_{\{010\}}(t) + \lambda_2 P_{\{100\}}(t) + 2\mu_1 P_{\{210\}}(t) + 2\mu_2 P_{\{120\}}(t) + \mu_3 P_{\{111\}}(t) \quad (71)$$

$$P'_{\{101\}}(t) = -(\lambda + \mu_1 + \mu_3) P_{\{101\}}(t) + \lambda_1 P_{\{001\}}(t) + \lambda_3 P_{\{100\}}(t) + 2\mu_1 P_{\{201\}}(t) + \mu_2 P_{\{111\}}(t) + 2\mu_3 P_{\{102\}}(t) \quad (72)$$

$$P'_{\{011\}}(t) = -(\lambda + \mu_2 + \mu_3) P_{\{011\}}(t) + \lambda_2 P_{\{001\}}(t) + \lambda_3 P_{\{010\}}(t) + \mu_1 P_{\{111\}}(t) + 2\mu_2 P_{\{021\}}(t) + 2\mu_3 P_{\{012\}}(t) \quad (73)$$

$$P'_{\{300\}}(t) = -(\lambda + 3\mu_1) P_{\{300\}}(t) + \lambda_1 P_{\{200\}}(t) + \mu_2 P_{\{310\}}(t) + \mu_3 P_{\{301\}}(t) \quad (74)$$

$$P'_{\{210\}}(t) = -(\lambda + 2\mu_1 + \mu_2) P_{\{210\}}(t) + \lambda_1 P_{\{110\}}(t) + \lambda_2 P_{\{200\}}(t) + 3\mu_1 P_{\{310\}}(t) + 2\mu_2 P_{\{220\}}(t) \quad (75)$$

$$P'_{\{201\}}(t) = -(\lambda + 2\mu_1 + \mu_3) P_{\{201\}}(t) + \lambda_1 P_{\{101\}}(t) + \lambda_3 P_{\{200\}}(t) + 3\mu_1 P_{\{301\}}(t) + \mu_2 P_{\{211\}}(t) + 2\mu_3 P_{\{202\}}(t) \quad (76)$$

$$P'_{\{120\}}(t) = -(\lambda + \mu_1 + 2\mu_2) P_{\{120\}}(t) + \lambda_1 P_{\{020\}}(t) + \lambda_2 P_{\{110\}}(t) + 2\mu_1 P_{\{220\}}(t) + \mu_3 P_{\{121\}}(t) \quad (77)$$

$$P'_{\{111\}}(t) \tag{78}$$
$$= -(\lambda + \mu_1 + \mu_2 + \mu_3)P_{\{111\}}(t) + \lambda_1 P_{\{011\}}(t) + \lambda_2 P_{\{101\}}(t) + \lambda_3 P_{\{110\}}(t) + 2\mu_1 P_{\{211\}}(t) + 2\mu_2 P_{\{121\}}(t) + 2\mu_3 P_{\{112\}}(t)$$

$$P'_{\{102\}}(t) = -(\lambda + \mu_1 + 2\mu_3)P_{\{102\}}(t) + \lambda_1 P_{\{002\}}(t) + \lambda_3 P_{\{101\}}(t) + 2\mu_1 P_{\{202\}}(t) + \mu_2 P_{\{112\}}(t) \tag{79}$$

$$P'_{\{021\}}(t) \tag{80}$$
$$= -(\lambda + 2\mu_2 + \mu_3)P_{\{021\}}(t) + \lambda_2 P_{\{011\}}(t) + (\lambda_2 + \lambda_3)P_{\{020\}}(t) + \mu_1 P_{\{121\}}(t) + 2\mu_3 P_{\{022\}}(t)$$

$$P'_{\{012\}}(t) \tag{81}$$
$$= -(\lambda + \mu_2 + 2\mu_3)P_{\{012\}}(t) + (\lambda_2 + \lambda_3)P_{\{002\}}(t) + \lambda_2 P_{\{011\}}(t) + \mu_1 P_{\{112\}}(t) + 2\mu_2 P_{\{122\}}(t)$$

$$P'_{\{310\}}(t) \tag{82}$$
$$= -(\lambda + 3\mu_1 + \mu_2)P_{\{310\}}(t) + \lambda_1 P_{\{210\}}(t) + \left(\frac{\lambda_1}{2} + \lambda_2\right)P_{\{300\}}(t) + 2\mu_2 P_{\{320\}}(t) + \mu_3 P_{\{311\}}(t)$$

$$P'_{\{301\}}(t) \tag{83}$$
$$= -(\lambda + 3\mu_1 + \mu_3)P_{\{301\}}(t) + \lambda_1 P_{\{201\}}(t) + \left(\frac{\lambda_1}{2} + \lambda_3\right)P_{\{300\}}(t) + \mu_2 P_{\{311\}}(t) + 2\mu_3 P_{\{302\}}(t)$$

$$P'_{\{220\}}(t) \tag{84}$$
$$= -(\lambda + 2\mu_1 + 2\mu_2)P_{\{220\}}(t) + \lambda_1 P_{\{120\}}(t) + \lambda_2 P_{\{210\}}(t) + 3\mu_1 P_{\{320\}}(t) + \mu_3 P_{\{221\}}(t)$$

$$P'_{\{211\}}(t) \tag{85}$$
$$= -(\lambda + 2\mu_1 + \mu_2 + \mu_3)P_{\{211\}}(t) + \lambda_1 P_{\{111\}}(t) + \lambda_2 P_{\{201\}}(t) + \lambda_3 P_{\{210\}}(t) + 3\mu_1 P_{\{311\}}(t) + 2\mu_2 P_{\{221\}}(t) + 2\mu_3 P_{\{212\}}(t)$$

$$P'_{\{202\}}(t) \tag{86}$$
$$= -(\lambda + 2\mu_1 + 2\mu_3)P_{\{202\}}(t) + \lambda_1 P_{\{102\}}(t) + \lambda_3 P_{\{201\}}(t) + 3\mu_1 P_{\{302\}}(t) + \mu_2 P_{\{212\}}(t)$$

$$P'_{\{121\}}(t) \tag{87}$$
$$= -(\lambda + \mu_1 + 2\mu_2 + \mu_3)P_{\{121\}}(t) + \lambda_1 P_{\{021\}}(t) + \lambda_2 P_{\{111\}}(t) + (\lambda_2 + \lambda_3)P_{\{120\}}(t) + 2\mu_1 P_{\{221\}}(t) + 2\mu_3 P_{\{122\}}(t)$$

$$P'_{\{112\}}(t) \tag{88}$$
$$= -(\lambda + \mu_1 + \mu_2 + 2\mu_3)P_{\{112\}}(t) + \lambda_1 P_{\{012\}}(t) + (\lambda_2 + \lambda_3)P_{\{102\}}(t) + \lambda_3 P_{\{111\}}(t) + 2\mu_1 P_{\{212\}}(t) + 2\mu_2 P_{\{122\}}(t)$$

$$P'_{\{022\}}(t) = -(\lambda + 2\mu_2 + 2\mu_3)P_{\{022\}}(t) + (\lambda_2 + \lambda_3)\left(P_{\{012\}}(t) + P_{\{021\}}(t)\right) + \mu_1 P_{\{122\}}(t) \tag{89}$$

$$P'_{\{320\}}(t) = -(\lambda + 3\mu_1 + 2\mu_2)P_{\{320\}}(t) + \lambda_1 P_{\{220\}}(t) + \left(\frac{\lambda_1}{2} + \lambda_2\right)P_{\{310\}}(t) + \mu_3 P_{\{321\}}(t) \tag{90}$$

$$P'_{\{311\}}(t) \tag{91}$$
$$= -(\lambda + 3\mu_1 + \mu_2 + \mu_3)P_{\{311\}}(t) + \lambda_1 P_{\{211\}}(t) + \left(\frac{\lambda_1}{2} + \lambda_2\right)P_{\{301\}}(t) + \left(\frac{\lambda_1}{2} + \lambda_3\right)P_{\{310\}}(t) + 2\mu_2 P_{\{321\}}(t) + 2\mu_3 P_{\{312\}}(t)$$

$$P'_{\{302\}}(t) = -(\lambda + 3\mu_1 + 2\mu_3)P_{\{302\}}(t) + \lambda_1 P_{\{202\}}(t) + \left(\frac{\lambda_1}{2} + \lambda_3\right)P_{\{301\}}(t) + \mu_2 P_{\{312\}}(t) \tag{92}$$

$$P'_{\{221\}}(t) \tag{93}$$
$$= -(\lambda + 2\mu_1 + 2\mu_2 + \mu_3)P_{\{221\}}(t) + \lambda_1 P_{\{121\}}(t) + \lambda_2 P_{\{211\}}(t)(\lambda_2 + \lambda_3)P_{\{220\}}(t) + 3\mu_1 P_{\{321\}}(t) + 2\mu_3 P_{\{222\}}(t)$$

$$P'_{\{212\}}(t) = -(\lambda + 2\mu_1 + \mu_2 + 2\mu_3)P_{\{212\}}(t) + \lambda_1 P_{\{112\}}(t) + (\lambda_2 + \lambda_3)P_{\{202\}}(t) + \lambda_3 P_{\{211\}}(t) + 3\mu_1 P_{\{312\}}(t) + 2\mu_2 P_{\{222\}}(t) \quad (94)$$

$$P'_{\{122\}}(t) = -(\lambda + \mu_1 + 2\mu_2 + 2\mu_3)P_{\{122\}}(t) + \lambda P_{\{022\}}(t) + (\lambda_2 + \lambda_3)\left(P_{\{112\}}(t) + P_{\{121\}}(t)\right) + 2\mu_1 P_{\{222\}}(t) \quad (95)$$

$$P'_{\{321\}}(t) = -(\lambda + 3\mu_1 + 2\mu_2 + \mu_3)P_{\{321\}}(t) + \lambda_1 P_{\{221\}}(t) + \left(\frac{\lambda_1}{2} + \lambda_2\right)P_{\{311\}}(t) + \lambda P_{\{320\}}(t) + 2\mu_3 P_{\{322\}}(t) \quad (96)$$

$$P'_{\{312\}}(t) = -(\lambda + 3\mu_1 + \mu_2 + 2\mu_3)P_{\{312\}}(t) + \lambda_1 P_{\{212\}}(t) + \lambda P_{\{302\}}(t) + \left(\frac{\lambda_1}{2} + \lambda_3\right)P_{\{311\}}(t) + 2\mu_2 P_{\{322\}}(t) \quad (97)$$

$$P'_{\{222\}}(t) = -(\lambda + 2\mu_1 + 2\mu_2 + 2\mu_3)P_{\{222\}}(t) + \lambda P_{\{122\}}(t) + (\lambda_2 + \lambda_3)\left(P_{\{212\}}(t) + P_{\{221\}}\right) + 3\mu_1 P_{\{322\}}(t) \quad (98)$$

$$P'_{\{322\}}(t) = -(\lambda + \mu)P_{\{322\}}(t) + \lambda\left(P_{\{321\}}(t) + P_{\{312\}}(t) + P_{\{222\}}(t)\right) + \mu P_{\{322|1\}}(t) \quad (99)$$

$$P'_{\{322|1\}}(t) = -(\lambda + \mu)P_{\{322|1\}}(t) + \lambda P_{\{322\}} + \mu P_{\{322|2\}}(t) \quad (100)$$

$$P'_{\{322|2\}}(t) = -(\lambda + \mu)P_{\{322|2\}}(t) + \lambda P_{\{322|1\}} + \mu P_{\{322|3\}}(t) \quad (101)$$

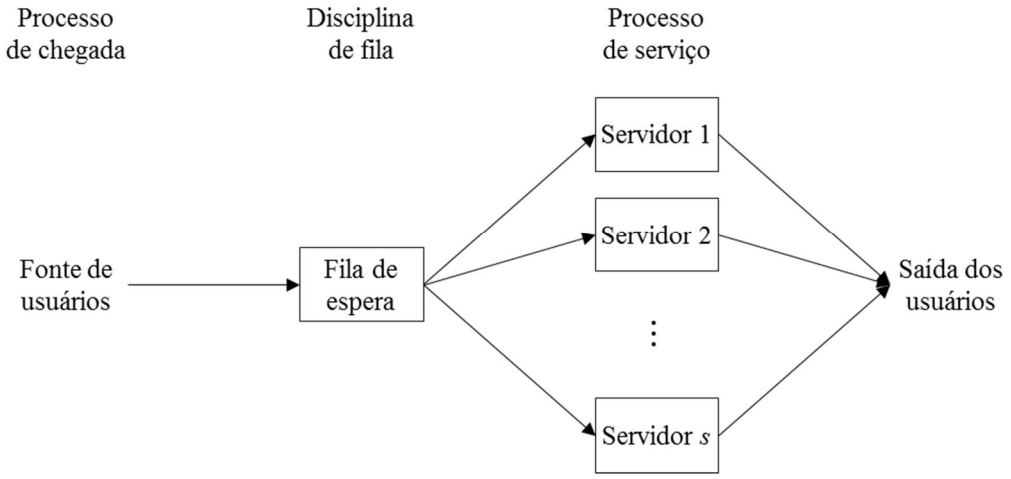
$$P'_{\{322|3\}}(t) = -(\lambda + \mu)P_{\{322|3\}}(t) + \lambda P_{\{322|2\}} + \mu P_{\{322|4\}}(t) \quad (102)$$

$$P'_{\{322|4\}}(t) = -\mu P_{\{322|4\}}(t) + \lambda P_{\{322|3\}} \quad (103)$$

ANEXO A

Genericamente, para descrever um sistema de filas, precisamos de informações a respeito de três elementos. O primeiro é o processo pelo qual usuários entram no sistema, chamado de processo de chegada. O segundo é a ordem em os usuários em fila são atendidos, chamada de disciplina de fila. O terceiro elemento é sobre o processo de serviço dos usuários, pelo qual saem do sistema (ARENALES *et al.*, 2015). A Figura 75 mostra um esquema resumindo a interação entre esses elementos.

Figura 75 – Esquema de funcionamento básico de um sistema de filas.



Para simplificar a análise dos modelos de fila, Kendall (1953) definiu uma notação para definir modelos de filas com um ou mais servidores idênticos em paralelo. A Notação é composta por três características *A*/*B*/*s*. Aqui, *A* e *B* são símbolos e *s* é uma constante inteira. *A* e *B* indicam a distribuição de probabilidade dos intervalos entre chegadas e dos tempos de serviço, respectivamente. A constante *s* indica o número de servidores idênticos trabalhando em paralelo no sistema (LARSON; ODONI, 2007).

A notação padrão para representar as distribuições de probabilidade de *A* e *B* são:

- *M*: distribuição exponencial (Chamada de Markoviana, sem memória);
- *D*: distribuição determinística;
- E_k: distribuição de Erlang com parâmetro de forma *k*;
- *GI*: distribuição genérica independente (para intervalos de tempo entre chegadas); e
- *G*: distribuição genérica de tempos de serviço.

Existem outras características possíveis de se apresentar nessas denominações como o tamanho da população fonte de usuários, a disciplina obedecida pela fila e a capacidade do

sistema. Apenas as três primeiras precisam ser sempre especificadas, caso as outras sejam omitidas, considera-se uma capacidade e população infinitas e uma disciplina de fila FCFS. Por exemplo, modelos *M*/*M*/*s* são modelos com intervalos entre chegadas e tempos de serviço que obedecem a uma distribuição exponencial e possuem *s* servidores atendendo em paralelo.

Em grande parte dos sistemas de filas, as chegadas de usuários são totalmente aleatórias. Isto significa que a chegada de um novo usuário independe do tempo transcorrido desde a última chegada ou saída (TAHA, 2008). Desta maneira, os intervalos entre chegadas podem ser descritos pela distribuição exponencial, com função de densidade de probabilidade, esperança e função acumulada definidas na Equação (104).

$$f(t) = \lambda e^{-\lambda t}, \qquad t > 0$$
$$E[t] = \frac{1}{\lambda} \tag{104}$$
$$F(t) = P(\text{intervalo de tempo entre chegadas} \leq t) = 1 - e^{-\lambda t}$$

Nesta definição, λ é a taxa por unidade de tempo pela qual chegam usuários ao sistema. A função acumulada representa a probabilidade de que o intervalo entre duas chegadas seja menor do que t.

A explicação para que a distribuição exponencial possa representar eventos totalmente aleatórios, resultado conhecido por falta de memória da distribuição exponencial, é apresentado na Equação (105).

$$P(t > T + S | t > S) = 1 - P(t \leq T) = P(t > T) = e^{-\lambda T}$$
$$P(t > T + S | t > S) = \frac{P(t > T + S, t > S)}{P(t > S)} = \frac{P(t > T + S)}{P(t > S)} = \frac{e^{-\lambda(T+S)}}{e^{-\lambda S}} \tag{105}$$
$$P(t > T + S | t > S) = e^{-\lambda T} = P(t > T) \blacksquare$$

Sem perda de generalidade, pode-se assumir que em $t = 0$ o sistema está vazio, $N(0) = 0$. Assim, o número de usuários no instante t é definido por $N(t)$ e a probabilidade de o sistema estar com n usuários é $P(N(t) = n)$. Dessa maneira, podemos definir uma probabilidade $p_0(\Delta t)$ como sendo a probabilidade de não haver nenhuma chegada durante o período $[t, t + \Delta t]$. Com isso, a Equação (106) mostra o cálculo de p_0.

$$p_0(\Delta t) = P(\text{intervalo de tempo entre chegadas} > \Delta t)$$
$$p_0(\Delta t) = 1 - P(\text{intervalo de tempo entre chegadas} \leq \Delta t) \tag{106}$$
$$p_0(\Delta t) = 1 - \left(1 - e^{-\lambda \Delta t}\right) = e^{-\lambda \Delta t}$$

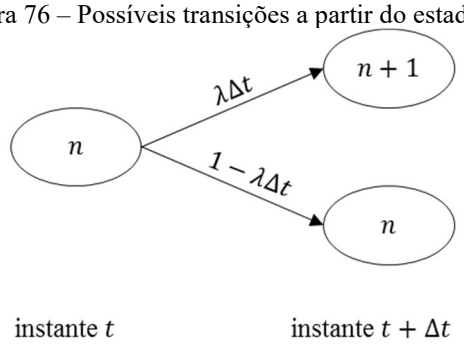
Isto é válido para qualquer instante de tempo t e intervalo de tempo Δt. Sendo que para um valor de Δt suficientemente pequeno temos por expansão em série de Taylor a Equação (107).

$$p_0(\Delta t) = 1 - \lambda\Delta t + \frac{(-\lambda\Delta t)^2}{2!} + \frac{(-\lambda\Delta t)^3}{3!} + \cdots = 1 - \lambda\Delta t + o(\Delta t) \tag{107}$$

Onde $o(\Delta t)$ tende a zero mais rapidamente do que Δt no limite $\Delta t \to 0$. Com isso, a probabilidade de haver duas chegadas ou mais de usuários também tende à zero. A probabilidade de haver uma chegada se torna o complemento da probabilidade de haver zero chegadas no intervalo Δt, como mostra a Equação (108). A Figura 76 ilustra essas possíveis transições a partir um estado qualquer n (ARENALES *et al.*, 2015).

$$\begin{aligned} p_0(\Delta t) &= 1 - \lambda\Delta t + o(\Delta t) \approx 1 - \lambda\Delta t \\ p_1(\Delta t) &= 1 - p_0(\Delta t) \approx \lambda\Delta t \end{aligned} \tag{108}$$

Figura 76 – Possíveis transições a partir do estado n.



Este resultado mostra que a probabilidade de ocorrer uma chegada no intervalo de tempo Δt é proporcional ao intervalo e a constante de proporcionalidade é dada pela taxa de chegada λ. Logo, a distribuição de probabilidade do número de usuários em um instante $t + \Delta t$ pode ser encontrada a partir da Equação (109). A probabilidade de o sistema estar com n usuários é resultado da seguinte combinação de eventos: (i) probabilidade de o sistema estar com n usuários no instante t e não haver novas chegadas no intervalo Δt; (ii) a probabilidade de o sistema estar com $n - 1$ usuários no instante t e um novo usuário chegar ao sistema durante Δt. Por outro lado, a probabilidade de o sistema estar com 0 usuários é resultado da probabilidade de o sistema ainda estar vazio no instante t e não ocorrer nenhuma chegada de usuário no intervalo Δt. Lembrando que estes produtos de probabilidades podem ser realizados visto que as chegadas são independentes (TAHA, 2008).

$$\begin{aligned} p_n(t + \Delta t) &= p_n(t)(1 - \lambda\Delta t) + p_{n-1}(t)\lambda\Delta t, \qquad n > 0 \\ p_0(t + \Delta t) &= p_0(\Delta t)(1 - \lambda\Delta t), \qquad n = 0 \end{aligned} \tag{109}$$

Rearranjando os termos e tomando o limite de $\Delta t \to 0$ obtém-se o sistema de equações diferenciais da Equação (110).

$$\lim_{\Delta t \to 0} \frac{P_n(t+\Delta t) - P_n(t)}{\Delta t} = \frac{dP_n(t)}{dt} = -\lambda P_n(t) + \lambda P_{n-1}(t), \qquad n > 0$$
$$\lim_{\Delta t \to 0} \frac{P_0(t+\Delta t) - P_0(t)}{\Delta t} = \frac{dP_n(t)}{dt} = -\lambda P_0(t), \qquad n = 0 \tag{110}$$

Mesmo envolvendo infinitas equações diferenciais, pode-se resolver o sistema admitindo que há $n = 0$ usuários no sistema em $t = 0$. O resultado disto é $P_0(t) = e^{-\lambda}$. E substituindo este resultado nas equações subsequentes obtém-se, por indução, a Equação (111). Este resultado corresponde à distribuição de Poisson para o processo de chegada de usuários, conhecido por modelo de nascimento (KLEINROCK, 1975).

$$P_n(t) = P_n(N(t) = n) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}, \qquad n \in \mathbb{N}_+ \tag{111}$$

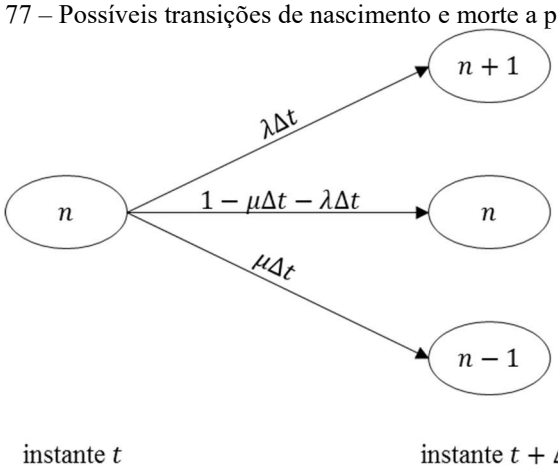
Para o caminho inverso, tem-se a saída de usuários, considerando que os tempos de serviço são exponencialmente distribuídos. O sistema começa com N usuários no instante 0 e os usuários saem do sistema seguindo uma taxa de serviço μ.

Utiliza-se os mesmos argumentos do processo de nascimento para desenvolver as equações diferenciais e obter a distribuição de probabilidade para o número de usuários que permanecem no sistema. O resultado é uma distribuição de Poisson truncada, conforme mostra a Equação (112). Para mais detalhes desta solução, sugere-se a leitura de Taha (2008). Este resultado representa o processo de saída do sistema, conhecido por modelo de morte puro.

$$P_n(t) = P_n(N(t) = n | N(0) = N) = \frac{(\mu t)^{N-n} e^{-\mu t}}{(N-n)!}, \qquad n \in \mathbb{N}_+ \tag{112}$$

Contudo, sistemas de filas costumam ser modelos de nascimento e morte, usuários entram e saem do sistema. Uma característica de grande valia desses modelos é que um sistema só pode mudar de uma situação para outra vizinha, como demonstrado ao longo das Equações (107) e (108) quando o intervalo de tempo analisado é suficientemente pequeno, $\Delta t \to 0$. Sendo assim, um sistema com $n > 0$ usuários no instante t pode sofrer os seguintes eventos neste intervalo de tempo: chegada de um novo usuário (nascimento), saída de um usuário (morte), ou permanência na situação anterior. A Figura 77 ilustra essas possibilidades de transições.

Figura 77 – Possíveis transições de nascimento e morte a partir de n.



Dessa forma, a distribuição de probabilidade do número de usuários no sistema no instante $t + \Delta t$ pode ser encontrada a partir da Equação (113). Agora, a probabilidade de o sistema estar no estado $n > 0$ é resultado da seguinte combinação de termos: (i) a probabilidade de o sistema estar com $n + 1$ usuários no sistema no instate t e ocorrer uma saída durante o intervalo Δt; (ii) a probabilidade de o sistema estar no estado n no instante t e não ocorrer nenhuma chegada ou saída ao longo do intervalo Δt; e (iii) a probabilidade de o sistema estar no estado $n - 1$ no instante t e ocorrer uma chegada ao longo de Δt (ARENALES *et al.*, 2015).

$$\begin{aligned} p_n(t + \Delta t) &= p_n(t)(1 - \lambda\Delta t - \mu\Delta t) + p_{n-1}(t)\lambda\Delta t + p_{n+1}(t)\mu\Delta t, \qquad n > 0 \\ p_0(t + \Delta t) &= p_0(\Delta t)(1 - \lambda\Delta t) + p_1(t)\mu\Delta t, \qquad n = 0 \end{aligned} \tag{113}$$

Rearranjando os termos e tomando o limite de $\Delta t \to 0$ pode-se encontrar o conjunto de equações diferenciais definido pela Equação (114).

$$\lim_{\Delta t \to 0} \frac{P_n(t + \Delta t) - P_n(t)}{\Delta t} = \frac{dP_n(t)}{dt} = -(\lambda + \mu)P_n(t) + \lambda P_{n-1}(t) + \mu P_{n+1}(t), \quad n > 0 \tag{114}$$

$$\lim_{\Delta t \to 0} \frac{P_0(t + \Delta t) - P_0(t)}{\Delta t} = \frac{dP_n(t)}{dt} = -\lambda P_0(t) + \mu P_1(t), \qquad n = 0$$

Esse sistema de equações diferenciais descreve o comportamento do sistema ao longo do tempo, sendo este chamado de comportamento transiente do sistema. (KLEINROCK, 1975). Como agora o modelo é de nascimento e morte, se a taxa média de chegada ao longo do dia for sempre menor do que a taxa total de serviço o sistema poderá atingir o equilíbrio, chamado estado estacionário. A Equação (115) mostra demonstra esta condição, onde essa precisa ser válida para qualquer instante t e intervalo de tempo τ (LARSON; ODONI, 2007).

$$\bar{\lambda} = \frac{1}{\tau}\int_0^{\tau} \lambda(t)dt < m\mu \tag{115}$$

O estado estacionário é atingido quando se considera valores de t suficientemente grandes. Sendo que a probabilidade de o sistema estar com n usuários passa a independer da situação inicial do sistema, como mostra a Equação (116).

$$\lim_{t\to\infty} P_n(t) = P_n \Rightarrow \frac{dP_n(t)}{dt} = 0 \tag{116}$$

O comportamento deixa de ser modelado pelas equações diferenciais e passa a ser modelado por um sistema de equações lineares. A Equação (117) ilustra as equações lineares de forma a mostrar que, em equilíbrio, a taxa média de saída do estado n, representada pelo lado esquerdo das equações, deve ser igual à taxa média de entrada no estado n, representada pelo lado direito das equações (KLEINROCK, 1975). Note que, para fins de generalização, cada estado possui agora uma taxa de chegada e uma taxa de serviço específica.

$$\begin{gathered}(\lambda_n + \mu_n)P_n = \lambda_{n-1}P_{n-1} + \mu_{n+1}P_{n+1}, \qquad n > 0 \\ \lambda_0 P_0 = \mu_1 P_1, \qquad n = 0\end{gathered} \tag{117}$$

Uma característica importante desse sistema de equações é a possibilidade de ser resolvido recursivamente em termos de P_0. A Equação (118) mostra uma solução geral para os estados em que n é maior do que 0. K_n é a razão entre as chegadas e saídas do sistema usada para fazer a recursão.

$$P_n = \frac{\lambda_0\lambda_1\lambda_2 \dots \lambda_{n-1}}{\mu_1\mu_2\mu_3 \dots \mu_n} P_0 = K_n P_0, \qquad n > 0 \tag{118}$$

Como a soma de todas as probabilidades de estado deve ser igual à 1, a Equação (119) mostra como calcular P_0 e assim poder obter as probabilidades de todos os estados do sistema. Observe que para isto ser válido, é preciso que $\sum_{n=1}^{\infty} K_n < \infty$, outra forma de verificar a Equação (115).

$$P_0 = \frac{1}{1 + \sum_{n=1}^{\infty} K_n} \tag{119}$$

ANEXO B

Os sistemas de equações diferenciais mostrados na Seções 3.2 e na Seção 5.3 são lineares, de primeira ordem e homogêneos, tais como a Equação (120).

$$\begin{aligned} P_1{}' &= a_{11}(t)P_1 + a_{12}(t)P_2 + \cdots + a_{1n}(t)P_n, \\ P_2{}' &= a_{21}(t)P_1 + a_{22}(t)P_2 + \cdots + a_{2n}(t)P_n, \\ &\vdots \\ P_n{}' &= a_{n1}(t)P_1 + a_{n2}(t)P_2 + \cdots + a_{nn}(t)P_n, \end{aligned} \tag{120}$$

É mais uma exceção do que uma regra que uma equação diferencial possa ser resolvida de forma exata e explícita por métodos demonstrativos (EDWARDS; PENNEY, 1995). Os modelos utilizados foram modelados pelo problema de valor inicial, como na Equação (121). Para calcular as medidas de desempenho, e todos as informações pertinentes aos modelos de filas não-estacionários, é adequado se ter uma tabela com os valores da solução desconhecida $y(x)$ ao longo do intervalo $[a, b]$.

$$\frac{dy}{dx} = f(x, y), \qquad y(a) = y_0 \tag{121}$$

Primeiramente, estas aproximações podem ser calculadas através do Método de Euler. Para descrever este método, primeiro escolhe-se um passo fixo $h > 0$ e consideramos os pontos da Equação (122):

$$x_0 = a, \qquad x_1, \qquad x_2, \ldots, \qquad x_n, \qquad \ldots, \text{ onde } x_{n+1} = x_n + h \tag{122}$$

Esses pontos são utilizados para calcular as aproximações y_n para os valores verdadeiros $y(x_n)$. Assim, busca-se as aproximações da Equação (123):

$$y_n \approx y(x_n) \tag{123}$$

Quando $x = x_0$ a taxa de variação de y em relação à x é $y' = f(x_0, y_0)$. Se y mantiver esta mesma taxa entre x_0 e $x_0 + h$, a variação em y será exatamente $h \cdot f(x_0, y_0)$. Desse modo, teríamos como a nova aproximação para o valor verdadeiro de $y(x_1)$, a Equação (124).

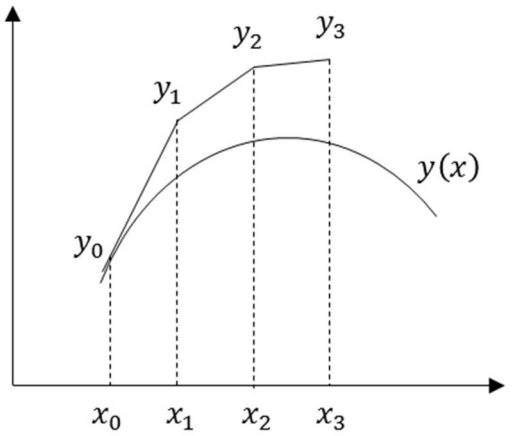
$$y_1 = y_0 + h \cdot f(x_0, y_0) \tag{124}$$

De maneira análoga, tendo alcançado o n-ésimo valor aproximado $y_n \approx y(x_n)$, tomamos como aproximação do valor verdadeiro, a Equação (125).

$$y_{n+1} = y_n + h \cdot f(x_n, y_n) \tag{125}$$

O Método de Euler consiste em aplicar, com passo h, a fórmula iterativa da Equação (125) para calcular aproximações sucessivas $y_1, y_2, y_3, \ldots$, para os valores verdadeiros $y(x_1), y(x_2), y(x_3), \ldots$, da solução exata $y = y(x)$ nos pontos $x_1, x_2, x_3, \ldots$, respectivamente. A Figura 78 ilustra a procedimento para este método.

Figura 78 – Ilustração do Método de Euler.



O erro do Método de Euler é da ordem h, isso permite obter qualquer nível de precisão desejado, escolhendo um valor para h suficientemente pequeno (EDWARDS; PENNEY, 1995).

Contudo, o Método de Euler é bastante assimétrico, já que assume a inclinação prevista no ponto x ao longo do intervalo $[x, x+h]$ como verdadeira. O Método de Euler Aprimorado (ou Aperfeiçoado) trabalha com uma previsão da inclinação média no intervalo $[x, x+h]$.

Suponha que após n passos de tamanho h, tenhamos calculado a aproximação y_n. Podemos usar o método de Euler para obter uma primeira estimativa, chamada agora de u_{n+1}, para o valor da solução em $x_{n+1} = x_n + h$, conforme a Equação (126).

$$u_{n+1} = y_n + h \cdot f(x_n, y_n) \tag{126}$$

Agora que $u_{n+1} \approx y(x_{n+1})$ foi calculado, pode-se tomar um k_{n+1}, calculado na Equação (127), como a predição de inclinação da curva solução $y = y(x)$ em $x = x_{n+1}$.

$$k_{n+1} = f(x_{n+1}, u_{n+1}) \tag{127}$$

O Método de Euler Aprimorado consiste em utilizar a média entre a inclinação k_n, já conhecida em $x = x_n$ e a predição k_{n+1} em $x = x_{n+1}$. Por simplicidade, referir-se-á aos valores de k_n e k_{n+1} por k_1 e k_2, respectivamente (BOYCE; DiPRIMA, 2001).

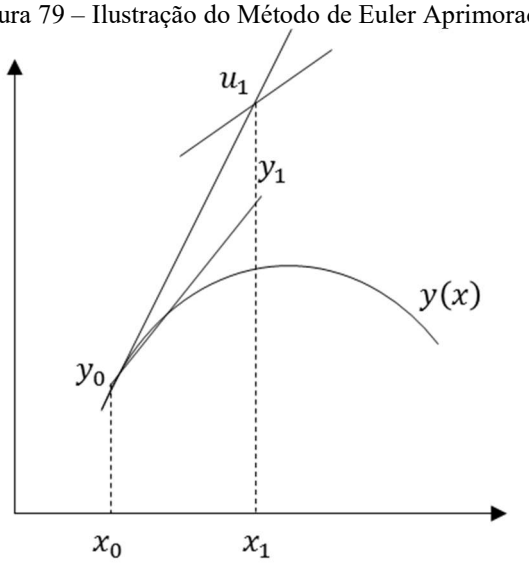
Sendo assim, o algoritmo do Método de Euler Aprimorado consiste em aplicar iterativamente as fórmulas da Equação (128) para calcular as aproximações sucessivas para os valores verdadeiros $y(x_1), y(x_2), y(x_3)$, ..., nos pontos x_1, x_2, x_3, ..., respectivamente.

$$\begin{aligned} k_1 &= f(x_n, y_n), \\ u_{n+1} &= y_n + h \cdot k_1, \\ k_2 &= f(x_{n+1}, u_{n+1}), \end{aligned} \tag{128}$$

$$y_{n+1} = y_n + \frac{h}{2}(k_1 + k_2)$$

O erro do Método de Euler Aprimorado é da ordem de grandeza h^2. A Figura 79 ilustra o Método de Euler Aprimorado.

Figura 79 – Ilustração do Método de Euler Aprimorado.



Por fim, o Método de Runge-Kutta é consideravelmente mais preciso do que o método de Euler Aprimorado. Considerando que deseja-se calcular $y_{n+1} \approx y(x_{n+1})$, então, pelo teorema fundamental do cálculo, tem-se a Equação (129) (EDWARDS; PENNEY, 1995).

$$y(x_{n+1}) - y(x_n) = \int_{x_n}^{x_{n+1}} y'(x)dx = \int_{x_n}^{x_n+h} y'(x)dx \tag{129}$$

Assim, a regra de Simpson para a integração numérica fornece a Equação (130).

$$y(x_{n+1}) - y(x_n) \approx \frac{h}{6}\left[y'(x_n) + 4y'\left(x_n + \frac{h}{2}\right) + y'(x_{n+1})\right] \tag{130}$$

Podendo, assim, definir y_{n+1} conforme a Equação (131).

$$y_{n+1} \approx y_n + \frac{h}{6}\left[y'(x_n) + 2y'\left(x_n + \frac{h}{2}\right) + 2y'\left(x_n + \frac{h}{2}\right) + y'(x_{n+1})\right] \tag{131}$$

Separou-se $4y'\left(x_n + \frac{h}{2}\right)$ em uma soma de termos para poder aproximar a inclinação $y'\left(x + \frac{h}{2}\right)$ no ponto $x = x_n + \frac{h}{2}$ do intervalo $[x_n, x_{n+1}]$ de dois modos diferentes.

As inclinações verdadeiras são substituídas pelas seguintes estimativas da Equação (132) (BOYCE; DiPRIMA, 2001):

- k_1: inclinação em x_n pelo método de Euler;

- k_2: estimativa da inclinação no ponto médio do intervalo $[x_n, x_{n+1}]$, utilizando o método de Euler para predizer a ordenada ali;
- k_3: este é o valor de Euler Aprimorado para a inclinação no ponto médio; e
- k_4: esta é a inclinação em $x = x_{n+1}$ pelo método de Euler, usando a inclinação aprimorada k_3 no ponto médio do segmento que terminava em $x = x_{n+1}$.

$$\begin{aligned} k_1 &= f(x_n, y_n) \\ k_2 &= f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}k_1\right) \\ k_3 &= f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}k_2\right) \\ k_4 &= f(x_{n+1}, y_n + hk_3) \end{aligned} \tag{132}$$

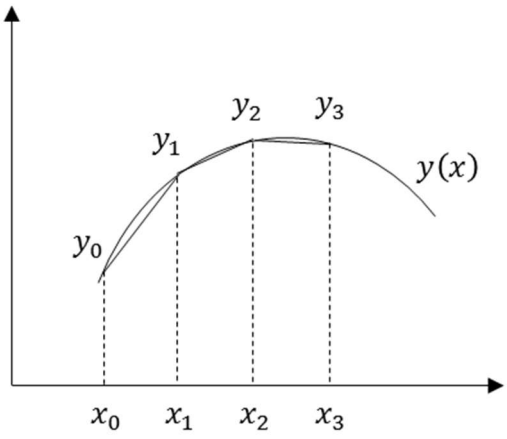
O algoritmo do Método de Runge-Kutta utiliza a fórmula iterativa da Equação (133) para calcular as aproximações sucessivas para os valores verdadeiros $y(x_1), y(x_2), y(x_3), ...$, nos pontos $x_1, x_2, x_3, ...$, respectivamente.

$$y_{n+1} = y_n + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4) \tag{133}$$

O erro para o método de Runge-Kutta é de ordem h^4, tendo uma precisão equivalente à aproximação com a fórmula de Taylor de quarto grau da Equação (134). No entanto, a aplicação do Método de Runge-Kutta é mais simples, já que usa apenas a função original $f(x, y)$, enquanto a fórmula de Taylor exige derivadas parciais de f de ordem superior (EDWARDS; PENNEY, 1995). A Figura 80 ilustra a precisão do método de Runge-Kutta.

$$y_{n+1} = y_n + y_n' h + \frac{y_n''}{2!}h^2 + \frac{y_n^{(3)}}{3!}h^3 + \frac{y_n^{(4)}}{4!}h^4 \tag{134}$$

Figura 80 – Ilustração do Método de Runge-Kutta.



ANEXO C

Muitas propriedades de uma série temporal podem ser captadas caso a série sofra uma decomposição dos seus componentes, como mostra a Equação (135) (EHLERS, 2005).

$$X_t = T_t + C_t + R_t \tag{135}$$

Em que:

- X_t é a série temporal propriamente dita;
- T_t é a componente da série que representa a tendência;
- C_t é a componente da série que representa a componente cíclica ou sazonal; e
- R_t é a componente aleatória, ou ruído.

Considera-se que a componente sazonal, C_t, se repete a cada intervalo fixo s, entendido como o tamanho do ciclo. Assim, variações periódicas podem ser captadas por esta componente (EHLERS, 2005).

A tendência pode ser entendida como uma mudança de longo prazo no nível médio da série. A Equação (136) mostra uma forma simples de definir uma série temporal com tendência.

$$X_t = \alpha + \beta t + \epsilon_t \tag{136}$$

Em que:

- α é uma constante que pode ser considerado um valor inicial para a série temporal;
- β é uma constante que representa o coeficiente de tendência da série temporal; e
- ϵ_t denota um erro aleatório com média 0.

A estimação de cada um dos parâmetros pode seguir um procedimento diferente. O software Minitab®, utilizado neste trabalho, segue os passos descritos a seguir para o modelo aditivo (MINITAB INC, 2010):

- 1º passo: os dados são suavizados utilizando uma média móvel centralizada com comprimento igual ao tamanho da componente sazonal. Caso a componente sazonal seja par, é necessária uma média móvel de duas etapas, a fim de sincronizar corretamente a média móvel;
- 2º passo: subtrai-se a média móvel dos dados para obter o que se denomina valores sazonais brutos;
- 3º passo: para períodos correspondentes dos ciclos sazonais, determina-se a mediana dos valores sazonais brutos. Por exemplo, em uma amostra de 60 meses (ou 5 anos) determina-se a mediana dos 5 valores sazonais correspondentes a cada mês do ano;
- 4º passo: ajusta-se as medianas dos valores brutos para que sua média seja zero. Essas medianas ajustadas constituem os índices sazonais;

- 5º passo: utiliza-se os índices sazonais para ajustar os dados sazonalmente; e
- 6º passo: ajusta-se uma linha de tendência aos dados ajustados sazonalmente usando regressão de mínimos quadrados. Lembrando que a tendência pode ser removida dos dados subtraindo o componente de tendência dos dados.

O método da decomposição também pode ser usado apenas para caracterização da série temporal, sem gerar previsões para os períodos futuros (EHLERS, 2005).

ANEXO D

A avaliação da hipótese de as chegadas de um sistema obedecerem a um processo de Poisson não-homogêneo (heterogêneo) requer que algumas considerações sejam feitas. Primeiramente, Brown *et al.* (2005) sugere um teste para um processo de Poisson cuja taxa de chegada varia lentamente ao longo do dia. Este teste não assume que as taxas de chegada dependem apenas do horário do dia e, portanto, seriam idênticos dia após dia. O teste também não requer o uso de variáveis adicionais para (tentar) estimar a taxa de chegada para uma data e hora específicas.

O primeiro passo para construir o teste é quebrar o dia em blocos de tempo relativamente pequenos de comprimento L, resultando em um total de I blocos. A suposição de blocos idênticos pode ser relaxada. Sugere-se que quanto menores forem as taxas de chegada, maiores devem ser seus intervalos. Também pode-se considerar um subconjunto de blocos, por exemplo, com dados de vários dias sucessivos no mesmo horário, ou blocos sucessivos. Seja T_{ij} o j-ésimo instante de chegada ordenado no i-ésimo bloco, $i = 1, \dots, I$. Então, $T_{i1} \leq \cdots \leq T_{iJ(i)}$, onde $J(i)$ indica o número total de chegadas no i-ésimo bloco. Assume-se $T_{i0} = 0$ e considera-se a transformação R_{ij} dada pela Equação (137).

$$R_{ij} = (J(i) + 1 - j)\left(-\log\left(\frac{L - T_{ij}}{L - T_{i,j-1}}\right)\right), \qquad j = 1, \dots, J(i) \tag{137}$$

Sob a hipótese nula de que os a taxa de chegada é constante dentro de cada intervalo de tempo, a transformação R_{ij} será um conjunto de variáveis exponenciais independentes e identicamente distribuídas.

Brown et al (2005) ainda sugere a consideração de uma propriedade do processo de Poisson, uniforme condicional (CU – *Conditional Uniform*, em inglês). De acordo com Kim e Whitt (2014), esta propriedade afirma que pode-se tomar, condicionalmente, ao número n de chegadas dentro de um intervalo $[0, L]$, os n horários de chegada ordenados, cada um dividido por L. Essas razões são distribuídas como a ordem estatística de n variáveis aleatórias independentes e identicamente distribuídas (i.i.d. – *independent identically distributed*, em inglês), cada uma uniformemente distribuída no intervalo $[0, 1]$. Assim, sob a hipótese nula de o processo seguir um processo de Poisson não-homogêneo, as n variáveis, com distribuição acumulada empírica F_n fica como dado pela Equação (138).

$$\bar{F}_n(t) \equiv \frac{1}{n}\sum_{j=1}^{n} 1_{\left\{\left(\frac{T_{ij}}{L}\right) \leq t\right\}}, \qquad 0 \leq t \leq 1 \tag{138}$$

Sendo que a o teste de Kolmogorov-Smirnov é feito pela Equação (139).

$$D_n \equiv \max_{0 \le t \le 1} |\bar{F}_n(t) - F(t)| \tag{139}$$

Onde $F(t) = t$. Para a hipótese nula ser aceita, é preciso que D_n seja comparado a um valor crítico. O valor crítico, δ é definido em função do nível de significância, α, e do número de observações, n, sendo $\delta \approx 1{,}36/\sqrt{n}$, para $n > 35$ e $\alpha = 0{,}05$. A hipótese nula é rejeitada quando $D_n > \delta$. Este teste é tratado por teste CU-KS.

Kim e Whitt (2014) ressalta que é importante lembrar que a propriedade CU elimina todos parâmetros de perturbação. Isso ajuda a testar processos de Poisson não-homogêneos com taxas de chegada constantes a cada período, pois possibilita a combinação de dados de intervalos separados com taxas diferentes. No entanto, é importante considerar que a taxa de chegada em cada intervalo pode ser aleatória, sendo que esta pode variar entre os dias, mesmo para os dias da mesma semana.

O teste de R_{ij}, apresentado por Brown *et al.* (2005), possui um maior poder com relação à processos alternativos com intervalos entre chegadas não-exponenciais. Contudo, Kim e Whitt (2014) encontrou uma transformação diferente para os dados ainda mais poderosa do que a R_{ij}. Esta transformação pode ser vista em Lewis (1965). O teste de Lewis é efetivo, porque foca nos intervalos entre chegadas.

O teste de Lewis propõe uma modificação diferente do teste CU-KS, explorando outra transformação. Começando com uma amostra U_j, $1 \le j \le n$, hipoteticamente sendo uniformemente distribuído em $[0, 1]$. Então considere $U_{(j)}$ como o j-ésimo menor desses, $1 \le j \le n$, de maneira que $U_{(1)} < \cdots < U_{(n)}$. Isso foi aplicado em Lewis (1965) com $U_{(j)} = T_j/t$ do teste CU-KS. Observa-se os sucessivos intervalos entre as observações ordenadas na Equação (140).

$$C_1 = U_{(1)}, \qquad C_j = U_{(j)} - U_{(j-1)}, \qquad 2 \le j \le n, \qquad \text{e } C_{n+1} = 1 - U_{(n)} \tag{140}$$

Considerando $C_{(j)}$ como o j-ésimo menor desses intervalos, $1 \le j \le n$, de forma que $0 < C_{(1)} < \cdots < C_{(n+1)} < 1$. Então considere Z_j, pela Equação (141), como versões escalonadas dos intervalos entre essas novas variáveis.

$$Z_j = (n + 2 - j)\big(C_{(j)} - C_{(j-1)}\big), \qquad 1 \le j \le n + 1, \qquad \text{com } C_{(0)} \equiv 0 \tag{141}$$

Sob a hipótese nula de o processo seguir um processo de Poisson, o vetor aleatório $(Z_1, \dots, Z_n)$ é distribuído da mesma maneira que o vetor aleatório $(C_1, \dots, C_n)$. Assim, considera-se o vetor de somas parciais associadas, S_k, dado pela Equação (142).

$$S_k \equiv Z_1 + \cdots + Z_k, \qquad 1 \le k \le n \tag{142}$$

Esse vetor $(S_1, \dots, S_n)$ possui a mesma distribuição que o vetor aleatório original $\left(U_{(1)}, \dots, U_{(n)}\right)$ de variáveis ordenadas aleatórias uniformes. De forma que o teste de Lewis consiste em verificar a distribuição acumulada empírica dada pela Equação (143), com $F(t) = t, 0 \leq t \leq 1$.

$$\bar{F}_n(t) \equiv \frac{1}{n}\sum_{j=1}^{n} 1_{\{S_k \leq t\}}, \qquad 0 \leq t \leq 1 \tag{143}$$

É importante ressaltar que Kim e Whitt (2014) sugere que, ao analisar um processo de Poisson não-homogêneo, o teste baseado na propriedade CU não deve ser descartado. Isso devido ao teste CU ser mais efetivo contra alternativas com intervalos de tempo exponenciais dependentes.

Algumas questões ainda permanecem, visto que se pode rejeitar, de forma inapropriada, a hipótese de seguir um processo de Poisson não-homogêneo devido à arredondamento dos dados. Uma forma de contornar isso é desarredondar os dados somando pequenas variáveis aleatórias uniformemente distribuídas. Caso o arredondamento não seja muito grosseiro, dados que não sigam um processo de Poisson continuarão não seguindo o processo após desarredondar.

Também pode-se rejeitar a hipótese de um processo seguir um processo de Poisson não-homogêneo por não se escolher os intervalos de tempo para análise de forma adequada. Isso porque os testes assumem que, dentro dos intervalos observados, a taxa de chegada é aproximadamente constante.

Por fim, a hipótese de um processo seguir um processo de Poisson não-homogêneo também pode ser rejeitada devido ao esforço para obter uma amostra maior. Isso quando se combina dados de diferentes dias, inapropriadamente. Quando não se leva em conta a possibilidade de haver variações entre os dias, mesmo os mesmos dias da semana.