UMA ANÁLISE DO SAMU DE OURO PRETO E MARIANA
utilizando filas dependentes do tempo

Arthur Piassi Dias de Castro

Editora Conhecimento Livre

Arthur Piassi Dias de Castro

Uma análise do SAMU de Ouro Preto e Mariana utilizando filas dependentes do tempo

1ª ed.

Piracanjuba
Editora Conhecimento Livre
2020

1ª ed.

Dados Internacionais de Catalogação na Publicação (CIP)

Castro, Arthur Piassi Dias de
C355u Uma análise do SAMU de Ouro Preto e Mariana utilizando filas dependentes do tempo. / Arthur Piassi Dias de Castro. – Piracanjuba-GO: Editora Conhecimento Livre, 2020.
49 f.: il.
ISBN: 978-65-80226-35-1
Modo de acesso: World Wide Web
Inclui bibliografia

1. Teoria de Filas. 2. Filas dependentes do tempo. 3. Sistema de Atendimento Emergencial. Arthur Piassi Dias de. I. Título.
CDU: 620

Sumário

FIGURA

TABELA

Agradecimentos

Aos meus pais, Rui e Denísia por sempre incentivarem meus estudos e me apoiarem nas minhas decisões.

Aos meus familiares por estarem presentes em todos momentos.

A Professora Lásara Rodrigues, pela orientação e confiança, e por não medir esforços para que esse trabalho fosse possível.

Aos demais professores da UFOP e COLTEC, pelos ensinamentos e aprendizado, sem vocês essa conquista não seria possível.

A todos meus amigos, vocês fizeram com que essa trajetória fosse muito mais prazerosa.

Aos moradores e ex-alunos da República Kaos, aprendi e cresci muito convivendo com vocês durante esses anos.

Arthur Piassi Dias de Castro

UMA ANÁLISE DO SAMU DE OURO PRETO E MARIANA UTILIZANDO FILAS DEPENDENTES DO TEMPO

Arthur Piassi Dias de Castro (Universidade Federal de Ouro Preto)

Resumo: O SAMU Inconfidentes é responsável pelo atendimento médico emergencial nos municípios de Ouro Preto e Mariana. O objetivo desse trabalho é modelar e analisar esse sistema por meio de Teoria de filas. Para tanto, são propostos modelos que visam representar as características desse sistema. Os modelos desenvolvidos utilizam-se da abordagem SIPP (*Stacionary Independent Period by Period*) e SIPP PRI (*Priority Stationary Independent Period by Period)*, onde assume-se o sistema como dependente do tempo com diferentes taxas de chegada de chamados ao longo dos períodos do dia, além de haver prioridade entre as classes de clientes. A variação na taxa de chegada pode resultar em picos de congestionamento e/ou ociosidade do sistema, de maneira que seja importante a identificação destes períodos e como seus resultados implicam no funcionamento geral dos atendimentos. Os resultados dos indicadores de desempenho foram calculados para todos os modelos e comparados aos observados na realidade com o objetivo de validar a modelagem de forma que ela possa ser utilizada para avaliar cenários e situações críticas dentro do sistema. Uma análise de custos também foi realizada com o objetivo de observar o impacto financeiro de se manter determinado cenário de atuação do sistema.

Palavras-chave: Teoria de Filas, Filas dependentes do tempo, SIPP, Sistema de Atendimento Emergencial.

1. INTRODUÇÃO

O Serviço de Atendimento Móvel de Urgência (SAMU 192) é um serviço de atendimento médico pré-hospitalar brasileiro. Esse serviço foi idealizado na França, em 1986, e nomeado Service d'Aide Médicale d'Urgence, daí a sigla "SAMU". No Brasil, esse serviço foi normalizado em 2004 pelo decreto presidencial nº 5.055, apesar de já estar em atuação em algumas cidades desde 1995.

Segundo Takeda et al. (2001), a função básica de um Serviço de Atendimento Móvel de Urgência (SAMU) é responder de forma organizada, a fim de evitar o uso excessivo de recursos, a toda situação de urgência que necessite de meios médicos, desde o primeiro contato telefônico até a liberação das vítimas ou seus encaminhamentos hospitalares. O sistema deve determinar e desencadear a resposta mais adequada para o caso, assegurar a disponibilidade dos meios hospitalares, determinar o tipo de transporte exigido e preparar o acolhimento dos pacientes.

Sistemas de saúde como o do SAMU são bastante sensíveis a atrasos por atenderem a várias ocorrências graves que envolvem risco de vida para quem está sendo atendido. Nesse caso, muitas das vezes, o atraso pode acarretar em agravamento do quadro clínico do paciente. Diversos autores, como Sanchez et al. (2010), McLay e Mayorga (2010) e Vukmir (2006) relatam a alta associação entre taxa de resposta rápida e sobrevivência dos pacientes, principalmente em ocorrências de alto nível de risco ou alta prioridade, como os tratados em sistemas de atendimento hospitalar e pré-hospitalar.

Segundo dados disponibilizados pela Secretaria Municipal de Saúde de Belo Horizonte, o tempo médio de resposta do SAMU na capital no ano de 2013 foi de 13 minutos, levando em consideração que o tempo máximo preconizado para 95% dos atendimentos é de 10 minutos, estes 3 minutos podem ser cruciais no socorro à vítima (JORNAL ESTADO DE MINAS, 2014).

De acordo com Green (2001), os atrasos são frutos de uma disparidade entre uma demanda e a capacidade de serviço disponível para atender esta demanda, ou seja, quando a taxa de realização do serviço é temporariamente menor do que a taxa em que chegam os pacientes. Normalmente isso acontece devido à uma variabilidade natural na demanda ou na própria taxa de serviço. Essa variabilidade e a interação entre a chegada e o serviço tornam a dinâmica desses sistemas bastante complexos. Portanto, para prever os níveis de recursos disponíveis para atender determinada performance podem ser usados modelo de filas.

Ainda de acordo com Green (2001), modelagens utilizando Teoria de Filas normalmente são relativamente simples de formular e, comparadas com metodologias de simulação, por exemplo,

requerem menos dados e menos tempo computacional. Isso faz com que seja mais fácil e barato utilizar-se de modelos de filas para obter resultados em sistemas que envolvem saúde. A simplicidade da utilização também torna mais direta a análise de possíveis cenários de variação na demanda e nos recursos disponíveis por meio dos indicadores de performance relacionados ao sistema.

O SAMU Inconfidentes atua nas cidades de Ouro Preto e Mariana, além de todos distritos pertencentes às duas cidades. Para isso, utiliza 4 ambulâncias, sendo 3 USBs (Unidades de Serviço Básico) e 1 USA (Unidade de Serviço Avançado), para realizar o atendimento à população. A região de Ouro Preto corresponde à uma área de 1.245,86 km² e uma população estimada de 73.994 pessoas, enquanto a de Mariana e seus distritos possui uma área de 1.194,21 km² e população de 60.142 pessoas, segundo o IBGE (2018).

1.1 OBJETIVO GERAL

O objetivo geral desse trabalho é modelar e analisar o funcionamento do SAMU na região dos Inconfidentes utilizando Teoria de Filas.

1.1.1 OBJETIVOS ESPECÍFICOS

Esse objetivo geral pode ser desmembrado nos seguintes objetivos específicos:

- Adaptar os modelos propostos ao sistema SAMU Ouro Preto e Mariana;
- Analisar as flutuações de chegada dos chamados ao longo do tempo;
- Calcular indicadores de desempenho para o sistema;
- Avaliar as variantes do SIPP PRI e adaptá-las ao sistema em questão;
- Simular cenários alternativos e compará-los com o cenário atual, avaliando as medidas de desempenho de cada cenário, além do custo relacionado.

1.2 JUSTIFICATIVA

Nos serviços médicos emergenciais, a rapidez na resposta ao chamado é um fator crucial para a sobrevida da vítima. Este tempo de resposta está diretamente relacionado a uma série de fatores, como o tráfego da região, o local de atendimento, o período do dia, o número de servidores disponíveis naquele momento, experiência e capacitação da equipe e a eficiência operacional do sistema (TAKEDA et al., 2001).

Em função disso, segundo Iannoni e Morabito (2006), esse tipo de sistema deve ser planejado com alto índice de ociosidade, para que se possa ter uma pequena probabilidade de congestionamento, ou seja, dos servidores que pertencem ao sistema estarem todos ocupados.

Embora desempenhe um papel fundamental para a qualidade de vida de toda a população, o SAMU assim como outros SAEs enfrenta restrições de orçamento, sendo impossível projetá-los com um número muito grande de servidores. Uma vez que, na visão econômica, ociosidade representa um gasto que pode ser eliminado, em muitos casos, é inviável atender à premissa proposta por Iannoni e Morabito (2006).

Devido a essas dificuldades orçamentárias e a necessidade de se organizar e reduzir o tempo de socorro à vítima que surge o propósito deste trabalho. Como os servidores móveis de atendimento são limitados para atender à população, este estudo visa entender como funciona o sistema de atendimento do SAMU e como os chamados se comportam ao longo do dia, de modo a identificar os horários mais críticos e com maior probabilidade de congestionamento do sistema, de modo a facilitar as tomadas de decisão dos seus gestores.

Do ponto de vista da modelagem, esse trabalho também se justifica uma vez que será utilizada a abordagem SIPP, especialmente o SIPP PRI, para avaliar as medidas de desempenho dos diferentes períodos do dia, além de comparar os seus resultados com modelos clássicos de filas e aproximações ou procedimentos heurísticos baseados nesses modelos.

1.3 ESTRUTURA DO TRABALHO

Este trabalho está organizado da seguinte maneira. A Seção 2 apresenta uma revisão de literatura sobre Teoria de Filas, além de se aprofundar em métodos de modelagem de sistemas de filas dependentes do tempo, apresentando exemplos e ressalvas sobre sua utilização. Já na Seção 3, é apresentado o Estudo de Caso, onde observa-se todos os parâmetros de funcionamento real do SAMU, ao passo que, a Seção 4 mostra os modelos propostos para a modelagem deste Estudo de Caso. A Seção 5 realiza um apanhado dos resultados obtidos pela modelagem e a Seção 6 propõe alguns cenários alternativos com o objetivo de analisar o comportamento do sistema, além de comparar os indicadores de desempenho obtidos para o sistema com os custos relacionados aos cenários. Finalmente, a Seção 7 apresenta uma conclusão dos resultados obtidos e sugere caminhos para trabalhos futuros envolvendo o mesmo objeto de estudo.

2. REFERENCIAL TEÓRICO

Esse capítulo apresenta os pressupostos teóricos que embasam este trabalho. Para um melhor entendimento dos temas, o capítulo se inicia com uma revisão geral dos sistemas de filas e filas dependentes do tempo para, em seguida, introduzir as abordagens SIPP e SIPP PRI.

2.1 SISTEMAS DE FILAS

A teoria de filas é um ramo da probabilidade estudada dentro da área de Pesquisa Operacional que consiste em analisar a formação de filas através de cálculos matemáticos que mensuram determinadas propriedades das filas. A Teoria de Filas proporciona a criação de modelos que permitem replicar as características de determinado sistema, além de simular situações hipotéticas dentro do mesmo.

Segundo Green et al. (2011), um modelo de filas é uma descrição matemática de um sistema de filas a partir do qual podem ser realizadas suposições específicas acerca da natureza probabilística do processo de chegada e serviço, o número e o tipo dos servidores, além da disciplina e organização da fila. Um sistema de filas básico é um sistema de serviços, onde “clientes” esperam algum tipo de serviço de um ou mais servidores que podem estar ou não disponíveis (GREEN et al., 2001).

Arenales et al. (2007) explicam que, em geral, os usuários desses sistemas se deslocam até os servidores para obter algum serviço, porém existem casos em que são os servidores que se deslocam para atender os usuários, como em casos de ambulâncias, bombeiros e viaturas de polícia. Nestes casos, os servidores são considerados móveis e as “filas” de usuários ficam dispersas, ao invés de concentradas fisicamente aguardando servidores fixos, como caixas de bancos, supermercados, etc.

A teoria de filas foi inicialmente motivada por aplicações em sistemas telefônicos e estuda as relações entre as demandas de um sistema e os atrasos sofridos pelos usuários do mesmo. A formação de filas ocorre se a demanda excede a capacidade do sistema de fornecer o serviço em um certo período. A teoria de filas auxilia no projeto e na operação dos sistemas para encontrar um balanceamento adequado entre os custos de oferecer serviços e os custos relativos aos atrasos sofridos pelos usuários (ARENALES et al., 2007).

Segundo Arenales et al. (2007), um sistema de filas é composto por três elementos básicos:

- A entrada ou processo de chegada de usuários no sistema;
- A disciplina da fila (ordem em que os usuários em fila são atendidos);

- A saída ou processo de serviço (atendimento).

A Figura 1 mostra como funciona este processo de maneira geral. Um conjunto de servidores presta atendimento à uma fonte de chegada ou população, sendo o sistema de filas composto por uma fila na qual o atendimento acontece de acordo com uma disciplina que determina qual é o próximo cliente a ser atendido.

O processo de chegada de usuários é descrito pelo intervalo de tempo entre chegadas sucessivas de usuários. Em geral, admite-se que o processo de chegada não varia ao longo do tempo e que o mesmo não é afetado pelo número de usuários do sistema. Pode-se, ainda, classificar os usuários como pertencentes a uma única classe ou agrupados em múltiplas classes diferentes.

A disciplina da fila corresponde a ordem em que os usuários são selecionados da fila para atendimento. A mais comum é primeiro a chegar; primeiro a ser servido. No caso de sistema de fila com prioridade, há o caso com interrupção ou sem interrupção, em que o processo de atendimento de um usuário de menor prioridade pode ser ou não interrompido em detrimento da chegada de um usuário de maior prioridade no sistema.

O processo de serviço é descrito pelo tempo de atendimento por usuário. O serviço pode ser realizado por um ou mais servidores, assim como a fila pode ser única ou não.

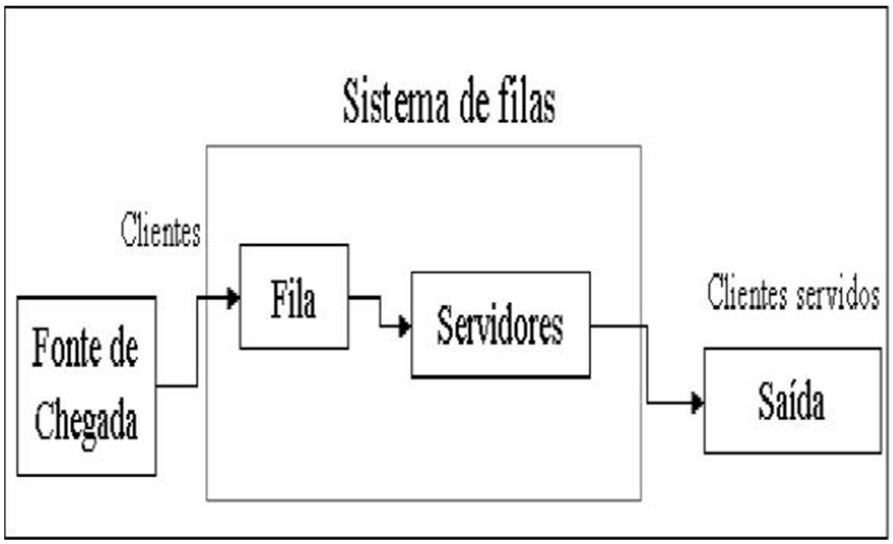

Figura 1 - Sistemas de Filas. Fonte: Silva (2008)

Arenales et al. (2007) classifica os sistemas de filas em quatro tipos distintos:

1. Fila única e um servidor;
2. Fila única e múltiplos servidores;

3. Múltiplas filas e múltiplos servidores em paralelo;
4. Fila única e múltiplos servidores em série.

2.2 SISTEMAS DE FILAS DEPENDENTES DO TEMPO

Green e Kolesar (1991) afirmam que a maioria dos sistemas de filas não apresentam taxas de chegada constantes e que é comum encontrar sistemas onde a demanda varia de acordo com o tempo de maneira cíclica, as quais podem apresentar variações diárias, semanais, mensais, etc. Isso pode ser visto em problemas envolvendo telecomunicações, bancos, sistemas policiais e emergenciais.

Os mesmos autores sugerem que qualquer sistema não estacionário pode ser tratado por meio de simulação ou utilizando-se de modelos de integração numérica, se forem Markovianos. Porém esses métodos apresentam, de maneira geral, um custo e tempo computacional altos. Portanto, vários métodos de aproximação utilizando filas foram desenvolvidos com o objetivo de simplificar e reduzir o custo de modelar esse tipo de sistema.

Segundo Schwarz et al. (2016), o estudo de sistemas de filas dependentes do tempo começou com Kolmogorov (1931) e desde então vem sendo utilizado em diferentes áreas de pesquisa, incluindo matemática, ciência da computação e pesquisa operacional. Os autores ainda ressaltam que utilizar entradas dependentes com o tempo geralmente gera um impacto grande na performance do sistema de filas, o que indica que é importante considerar essa variável ao modelar e controlar tais sistemas.

Schwarz et al. (2016) desenvolveram um esquema de classificação de sistemas de filas dependentes do tempo. O objetivo do esquema é classificar as diferentes abordagens encontradas na literatura, avaliar a performance dessas abordagens e estabelecer um paralelo entre os diferentes modelos. Os autores identificaram três principais categorias de abordagens:

1. Soluções numéricas ou analíticas: esse tipo de abordagem utiliza sistemas de equações para descrever o comportamento dependente do tempo de sistema de filas.
2. Abordagens que dividem o tempo total em intervalos menores: essa técnica modela cada período de tempo separadamente e trata os períodos como independentes, conectados ou transientes.
3. Abordagens baseadas em modificações nas características do sistema: utilizam-se de modificações nas propriedades do sistema, como mudanças no número de servidores ou nas

características do serviço para aproximar o modelo do sistema real sem a necessidade de dividir os períodos.

Dentro das abordagens que dividem o tempo total em intervalos menores e tratam os períodos como independentes, Schwarz et al. (2016) citam o SSA (Simple Stacionary Aproximation) o PSA (Pointwise Stacionary Aproximation) e o SIPP (Stationary Independent Period by Period). Essas abordagens diferem entre si quanto a duração do período e determinação dos parâmetros de entrada do sistema. No SSA, os parâmetros são uma média utilizada para todo o horizonte de tempo, o PSA, determina um período l=0 para se utilizar de valores paramétricos instantâneos, realizando suas análises de desempenho, já o SIPP modela cada período como independente dos outros, considerando uma taxa de chegada estacionária para cada período.

2.3 STATIONARY INDEPENDENT PERIOD BY PERIOD (SIPP)

O SIPP é uma abordagem de resolução de um sistema de filas que utiliza-se do artifício de separar o horizonte de tempo *T* em intervalos pré-definidos $[a_i, b_i]$, *(i =1, 2,...,I)* de maneira que sejam permitidas mudança em parâmetros de entrada do sistema em cada período. Dessa forma, pode-se aplicar um modelo estacionário de filas separadamente para cada um destes intervalos. Segundo Green (2011), a abordagem utilizada mais frequentemente na literatura é aplicar um modelo M/M/c para cada intervalo de tempo, obtendo assim um número de servidores mínimos para atender ao período em questão. O SIPP vêm sendo usado por muitos autores e apresenta bons resultados para resolver sistemas de filas M(t)/M/c(t) e M(t)/M/c (ver, por exemplo, Green et al. (2001) e Stolletz (2008, 2011)), com algumas ressalvas que serão abordadas posteriormente.

Green et al. (2001) estabelece três requisitos para a aplicação do SIPP:

1. O tempo entre a chegada e o atendimento (delay times) devem ser estatisticamente independentes entre si em intervalos consecutivos;
2. Em cada intervalo, o sistema entra em estado estacionário;
3. A taxa de chegada é constante dentro de cada intervalo de tempo.
4. É importante ressaltar que a escolha da duração de cada período pode influenciar diretamente no desempenho do método, bem como na sua aderência ao modelo de filas. Segundo Green (2011), quanto menor a duração do intervalo de tempo definido, é mais provável que o sistema adira ao SIPP. Os autores utilizaram períodos de ¼ de hora e 2 horas e os resultados mostraram

que 79% e 28% dos cenários são confiáveis para o SIPP, respectivamente. Outro resultado interessante é que a confiabilidade do SIPP tende a se reduzir quando a taxa de chegada aumenta ou quando ela apresenta maior variação de um período para outro. Esses resultados sugerem que o SIPP seja confiável com taxas de utilização baixas e curtos intervalos de tempo.

Stolletz (2011) mostra que a modelagem por meio do SIPP pode superestimar a utilização de recursos e tamanho da fila em sistemas onde se observa alta dependência entre as taxas de chegada de períodos consecutivos. Para estes casos, sugere-se a utilização da abordagem PSA que considera uma ligação entre as taxas de diferentes períodos. Já Stolletz (2013) utiliza uma variante do SIPP para resolver ou minimizar esse problema, o *lagged SIPP,* onde uma parte da taxa de chegada do período *i-1* é incorporada ao período *i*, considerando que há um atraso entre a chegada de uma ocorrência e a realização do serviço.

2.3.1 PRIORITY STATIONARY INDEPENDENT PERIOD BY PERIOD (SIPP PRI)

Em alguns casos, não se pode aplicar a abordagem SIPP devido às características do sistema em estudo. Um exemplo se dá quando existem diferentes classes de chamados em um sistema de fila e existe prioridade entre essas classes. Nesse caso, pode-se usar o chamado SIPP PRI, em que o "PRI" está associado ao conceito de prioridade.

Segundo Vile et al. (2016), o SIPP PRI é uma aproximação que estende o SIPP para tratar fila com prioridades, e pode considerar duas ou mais classes.

Como visto em Vile et al. (2015), sistemas de filas com duas classes distintas podem ser tratados dividindo-as em alta prioridade e baixa prioridade, obedecendo a disciplina de prioridade não-preemptiva (NPRP). Nesse caso, um chamado de baixa prioridade só pode ser atendido se não houverem chamados de alta prioridade na fila e nenhum atendimento pode ser interrompido em detrimento de outra chamada.

Vile et al. (2016) afirmam que o SIPP PRI é eficiente para estimar o número de equipes necessárias para atendimento, porém propõem algumas variantes na abordagem desenvolvidas para o caso específico modelado para que sejam alcançadas estimativas mais confiáveis. Dentre essas variantes, pode-se citar:

- Lag Avg Pri: Considera o atraso (*lag*) entre as chegadas e o congestionamento do sistema. Utilizado para corrigir uma possível inconsistência entre o pico de chamados e o pico de

congestionamento do sistema, que ocorre em função do atraso entre o chamado e o atendimento em si. Essa característica é problemática, pois se afasta da premissa de independência entre os períodos e, além disso, é possível que os indicadores de desempenho obtidos se mostrem falsos para determinado período de pico;

- SIPP Mix Pri: Essa variante utiliza uma taxa mínima de chegada em períodos em que a taxa de chegada estiver abaixo da média. O objetivo é tentar fixar um número mínimo de equipes para atuar no atendimento do sistema, para evitar falta de recursos em uma eventual demanda incomum dentro de um período;
- Adaptative SIPP Pri: Quando a taxa de chegada do período em questão difere muito do período anterior, a taxa utilizada é calculada pela média dos dois períodos, com o objetivo de diminuir a amplitude associada aos intervalos de tempo, o que, segundo os autores, costuma ser um problema em sistemas modelados utilizando-se do SIPP.

Outras variantes também podem ser propostas para tratar casos específicos. Além disso, as variantes já propostas podem ser utilizadas apenas com alteração de parâmetros para se adequarem ao sistema a ser modelado.

É importante notar que a confiabilidade do SIPP PRI é análoga aos casos de modelagem com SIPP, ou seja, o método se mostra confiável quando há taxas de utilização baixas, variabilidade baixa entre as taxas de chegada de períodos consecutivos e o tempo é dividido em intervalos curtos, como mostram Vile et al. (2016) ao apresentarem um estudo de caso, onde o SIPP PRI e suas variantes se mostram confiáveis apenas quando a amplitude da taxa de chegada entre os períodos estava entre 0 e 0,5 e os intervalos de tempo entre 0,25 e 0,5 horas. Essas condições se mostram bastante específicas e difíceis de modelar. Porém, é importante notar que o exemplo específico tratado no artigo apresenta um sistema relativamente congestionado, tornando difícil admitir a independência entre períodos. Portanto, é razoável inferir que em sistemas com taxas de utilização mais baixas, possa-se permitir parâmetros mais flexíveis de amplitude e divisão de intervalos de tempo. Essa suposição, contudo, precisa ser confirmada utilizando outros estudos e/ou por meio de simulação.

Vile et al. (2015) apresentam uma comparação entre os resultados de modelagens envolvendo sistemas de filas M(t)/M/s(t)/NPRP. Os autores utilizam um estudo de caso ilustrativo onde comparam o dimensionamento de recursos obtidos da modelagem utilizando SIPP PRI comparada com o EULER PRI, método numérico de resolução de sistemas de filas. Os resultados das aproximações do SIPP PRI

se mostram dependentes de condições específicas da modelagem. É ressaltado que, quando os tempos de serviço são longos e quando há uma variabilidade grande nas taxas de chegada em períodos próximos, os resultados se mostram pouco assertivos, casos onde recomenda-se a utilização do método exato.

Outros autores, como Ingolfsson et al. (2007), Vile et al. (2017) e Izady (2010, 2012), também realizam comparações entre os métodos exatos e métodos aproximados como o SIPP PRI em casos envolvendo sistemas M(t)/M/c. Além do método EULER citado anteriormente, também são utilizados o método de randomização e o método de Range-Kutta. Em Vile et al. (2017), os autores destrincham tais métodos para depois propor uma modelagem utilizando-se dos conceitos do SIPP PRI. Neste caso, define-se um tempo x, de acordo com as características do sistema, de forma que o tempo médio de espera em fila não possa ser maior que o mesmo. Esta estratégia baseia-se no conceito de dimensionamento de equipes (staffing), portanto o modelo é executado para várias quantidades de equipes com o objetivo de se obter o limite inferior que permita o sistema atuar dentro do tempo máximo de espera em fila.

3. ESTUDO DE CASO

Este capítulo aborda de forma geral o sistema objeto de estudo do trabalho. Na seção 3.1 são apresentadas as características do mesmo e realizadas análises estatísticas. Já na seção 3.2, são propostos modelos matemáticos que visam representar o sistema real, além da seção 3.2, que apresenta os resultados obtidos de acordo com a modelagem. Finalmente, a seção 3.3, propõe uma análise econômica, onde são comparados cenários alternativos para o SAMU, utilizando-se também dos indicadores calculados pelos modelos propostos.

3.1 O SAMU DE OURO PRETO E MARIANA

Nesta seção, o sistema SAMU é apresentado, antes de relatar como foi feita a coleta e análise de dados, partindo para a análise da distribuição temporal, análise de homogeneidade dos servidores e testes de aderência.

O SAMU de Ouro Preto e Mariana atua de forma integrada compartilhando os recursos dos municípios e utilizando políticas de despacho específicas dependendo do tipo e localização do chamado. A chegada das ocorrências ocorre pela discagem ao número 192, que encaminha os chamados para uma central localizada na cidade de Belo Horizonte. Durante o atendimento telefônico, a equipe decide

acionar ou não uma equipe de resgate. Além disso, é registrado o local da ocorrência e acionada a base mais próxima do mesmo. Nesse atendimento, os chamados também são classificados em:

- Chamados A - Prioridade alta: Ocorrências graves que sugerem um potencial risco de vida do paciente, que incluem traumas graves ou quadros clínicos relacionados a doenças com alta probabilidade de óbito ou que requerem cuidados especializados.
- Chamados B - Prioridade baixa: Ocorrências relacionadas à traumas menos graves, pequenos e médios ferimentos ou quadros clínicos estáveis.

As ambulâncias utilizadas pelo sistema incluem uma Unidade de Serviço Avançado (USA) e três Unidades de Serviço Básico (USB). As USBs contam com equipes formadas por motorista e um auxiliar ou técnico de enfermagem, enquanto a USA atua com os mesmos profissionais, além de um médico especializado e é melhor equipada, contando com aparelhos de ressuscitação. As ambulâncias USBs geralmente atendem à chamados B, porém, em casos da USA estar ocupada podem ser acionadas para chamados de prioridade alta. A USA atende somente os chamados de prioridade A ou transferências. As transferências são realizadas de maneira arbitrária pelos gerentes do sistema, porém afetam o desempenho e as taxas de utilização dos recursos, portanto têm que ser consideradas na análise.

Atualmente, as USB, quando disponíveis, estão localizadas nas UPA (Unidades de Pronto Atendimento) de Mariana, Ouro Preto e Cachoeira do Campo e a USA na UPA de Ouro Preto. Ressalva-se que até setembro de 2017, a USB localizada na UPA de Cachoeira do Campo estava localizada na UPA de Ouro Preto.

3.1.2 COLETA E ANÁLISE DE DADOS

A coleta e análise de dados foi autorizada pelas Secretarias Municipais de Saúde de Ouro Preto e Mariana e pelos coordenadores do SAMU de ambos os municípios. A coleta de dados foi realizada nas sedes do SAMU, localizadas nas UPAs de Ouro Preto e Mariana. Os dados coletados correspondem aos atendimentos realizados pelas quatro ambulâncias durante o ano de 2017. Esses dados foram essencialmente coletados em prontuários de atendimentos.

No total, foram identificados 4269 prontuários, dos quais 4145 foram consideradas válidos. Desse total, 1560 correspondiam à cidade de Mariana e seus distritos, enquanto 2331 à Ouro Preto e seus distritos, além de 147 ocorrências em rodovias, 14 em outras cidades e 93 consideradas como indeterminadas.

Um total de 4145, ou 97,10%, dos prontuários analisados, possuem a informação da hora de chegada do chamado, através da data e hora da chegada dos chamados, foi possível calcular o intervalo entre chegadas de chamados. O sistema apresenta um intervalo médio entre chegada de chamados de 126,81 minutos, pouco mais de 2 horas, o que significa que, em média, a cada intervalo de tempo desse, um chamado ocorre no sistema. O desvio padrão dos dados de chegada foi da mesma ordem de grandeza da média, 147,91 minutos.

Outro fator importante que se procurou analisar no sistema foi o tempo de serviço total, que inicia-se com a saída do servidor para realizar o atendimento e termina com a liberação do servidor, inclui: o tempo de viagem até o local de atendimento, o tempo em cena, o tempo de deslocamento do paciente até o hospital (em caso de necessidade de remoção) e o tempo de deslocamento do local de atendimento ou hospital de volta à base.

Os dados considerados válidos para essa análise foram 412 para a USA0009, correspondendo a 87,40% do total de atendimentos válidos. No caso das USBs, as amostras de dados válidos representaram 79,52% para a USB 7555, 74,66% para a USB 7556 e 95,40% para a USB 7557.

3.1.3 ANÁLISE DA DISTRIBUIÇÃO TEMPORAL

Visto que as chegadas de chamados não acontecem de maneira constante, buscou-se encontrar padrões no comportamento das mesmas, analisando a existência de tendência nas ocorrências em meses, semanas, dias ou horários específicos do dia.

Em relação a distribuição mensal, semanal e diária, foram aplicados testes de hipóteses utilizando o método de Tukey para avaliar a homogeneidade das chegadas ao longo desses períodos. Esses testes de hipótese constataram com 95% de confiança que não há distinção significativa entre a taxa de chegada ao longo desses períodos, ou seja, elas podem ser consideradas homogêneas.

Em seguida, a distribuição dos chamados ao longo do dia foi analisada. Para isso, dividiu-se o dia em períodos com o objetivo de entender o comportamento das chegadas ao longo do dia.

A Figura 2 mostra a divisão dos chamados em períodos de 4 horas cada. Os resultados mostram que os dois primeiros períodos do dia, de 0h às 8h apresentam uma quantidade de ocorrências bastante inferior aos períodos seguintes, com uma quantidade de chamados mais de duas vezes maior. O teste de Tukey confirmou a suposição de heterogeneidade entre os períodos, de modo que, com um nível de significância de 95%, os mesmos foram considerados distintos.

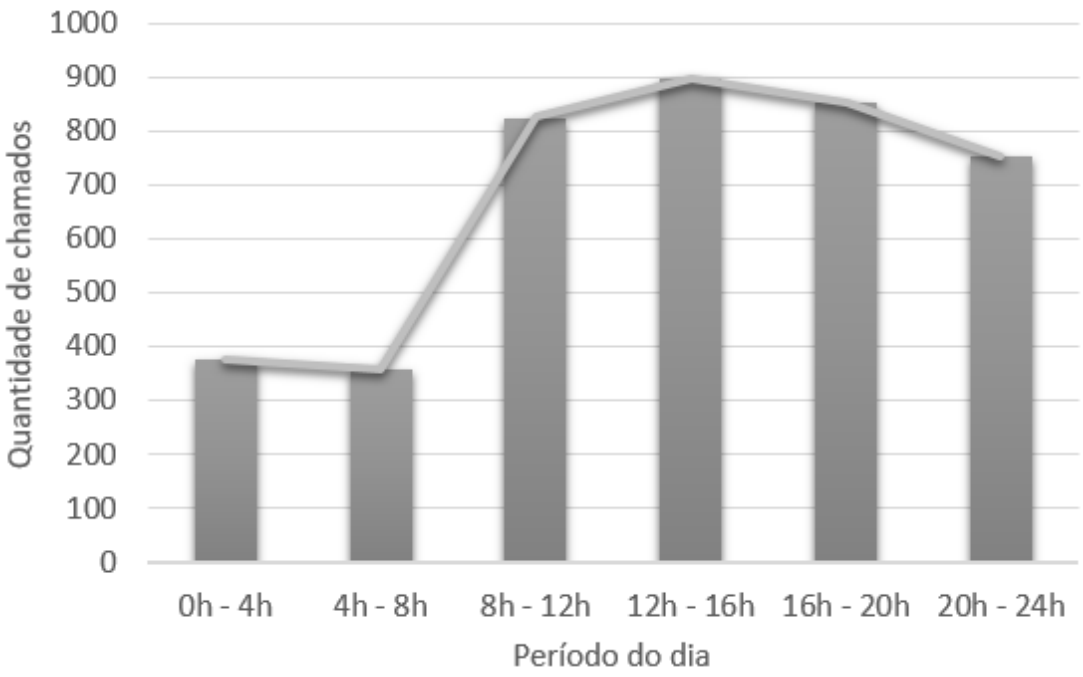

Figura 2 - Quantidade de chamados em períodos de 4 horas

A Figura 3 apresenta a distribuição temporal ao longo do dia ao utilizar intervalos de 1 hora. Pode-se observar significativa variação nos valores encontrados, por exemplo, o valor máximo de 241 ocorrências no período de 14h às 15h, valor 415% maior do que os 58 chamados observados no intervalo de 5h às 6h da manhã. O teste de Tukey confirmou, com uma precisão de 95%, que a quantidade de chamados, ao utilizar períodos de 1 hora, não pode ser considerada homogênea.

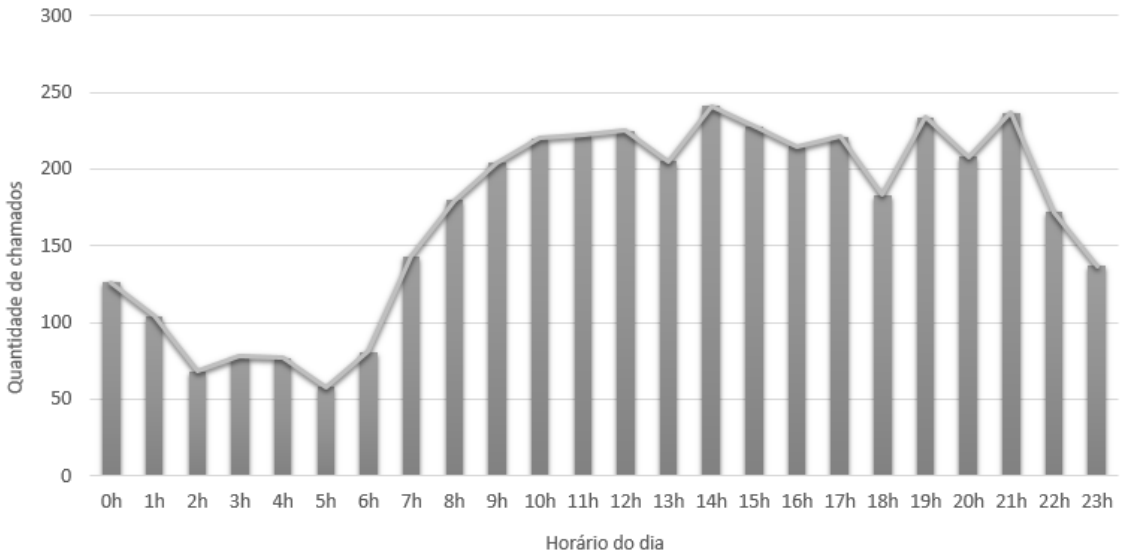

Figura 3 - Distribuição temporal por hora do dia

Pode-se observar também na Figura 3 que, a partir das 23 horas, o número de chamados começa a diminuir, se mantendo abaixo da média (de 149,46 atendimentos horários no ano) até as 8 horas da

manhã. Por outro lado, entre 9 e 22hs, as ocorrências se mantêm superior ao valor médio encontrado. Esses resultados sugerem que os horários de maior congestionamento do sistema são no fim da manhã, passando por toda a tarde e início da noite.

3.1.4 ANÁLISE DE SERVIÇO E HOMOGENEIDADE DOS SERVIDORES

Em relação ao atendimento, há de se ponderar que as ambulâncias estão localizadas em bases distintas, realizam atendimentos à diferentes distâncias e tipos de chamados. Essa situação sugere a necessidade de avaliar a homogeneidade dos servidores. Vale destacar que essa homogeneidade não está necessariamente associada a características do veículo e da equipa alocada, uma vez que pode ser relacionada ao tipo de chamado, localização da base e dos chamados.

Portanto, realizou-se a análise do tempo em cena, ou seja, o tempo em que a ambulância permanece no local de atendimento, e do tempo de serviço total, que corresponde ao intervalo de tempo desde que a mesma sai da base até o momento em que a ambulância encerra o seu atendimento, retornando para a base. O objetivo desse teste é avaliar se os servidores podem ser considerados homogêneos ao considerar o tempo em cena e o tempo de serviço total.

Ao analisar o tempo em cena dos chamados por ambulâncias, com a ajuda do teste de Tukey, constatou-se que as USBs podem ser consideradas homogêneas, ou seja, com um nível de significância de 95%, seus tempos em cena ou tempos de atendimento não podem ser considerados distintos. Já a USA apresentou um tempo consideravelmente maior, sendo classificada como heterogênea em relação as demais.

Já na análise do tempo de serviço total das ambulâncias, os resultados do teste de Tukey com nível de significância de 95% apontam que apenas as USBs de Ouro Preto (USB 7555 e USB7556) podem ser consideradas homogêneas entre si. Neste caso, a USB 7557, localizada na cidade de Mariana-MG, não foi agrupada com as USBs de Ouro Preto/Cachoeira do Campo. No caso, o tempo médio de atendimentos foi de 64,27 minutos, valor 15,27% menor que o tempo médio das demais unidades de serviço básico. A USA, por sua vez, apresenta tempo de serviço maior do que as demais, com uma média de 118,31 minutos, valor 68,43% maior que a média das UBSs. O tempo em trânsito da USA também é consideravelmente superior ao das USBs, isso se deve ao fato da USA realizar proporcionalmente mais transferências do que as USBs para hospitais de referência em Belo Horizonte, além de realizar atendimento em todas as regiões analisadas.

Em relação ao tempo em trânsito da USB 7557, os resultados confirmam as expectativas uma vez que a mesma realiza a grande maioria de seus atendimentos na região urbana de Mariana, resultando em um tempo de deslocamento menor que as de Ouro Preto. Visto que o tempo total de atendimento é composto pelo tempo em cena somado ao tempo de deslocamento da ambulância, estes resultados sugerem que o deslocamento é fator de influência no processo de serviço do sistema.

3.1.5 TESTES DE ADERÊNCIA

Foram também realizados testes de hipótese para verificar a aderência dos intervalos entre chegadas em cada período e do tempo de serviço total à distribuição exponencial.

A Tabela 1 mostra que, na divisão de tempo de 4 horas, todos os intervalos aderiram a distribuição com uma significância de 92%. Já no caso da divisão do dia em períodos de 1 hora, não se pode garantir a aderência à distribuição exponencial, como pode ser visto na Tabela 2. Estes resultados podem ser atribuídos a uma divisão muito grande do montante inicial de dados, o que causa um número amostral pequeno em cada período, resultando em flutuações naturais que divergem do que pode ser visto em uma distribuição exponencial. A primeira coluna das Tabelas 1 e 2 se refere ao período do dia em que foi feita a análise de chegadas. A segunda se refere à quantidade de chamados neste período. As colunas 3 e 4 mostram a média e o desvio padrão do intervalo entre os chamados, enquanto a última diz respeito ao p-valor encontrado no teste de aderência.

Tabela 1 - Aderência de cada período à distribuição exponencial (4h)

Período	Quantidade de chamados	Média de tempo entre chamados (minutos)	Desvio Padrão (minutos)	Aderência a Distribuição Exponencial (p-valor)
0-4 horas	364	227,2	187,2	<0,003
4-8 horas	330	148,9	120,1	<0,003
8-12 horas	813	103,0	104,2	0,080
12-16 horas	842	103,8	104,9	0,046
16-20 horas	781	106,3	112,3	0,041
20-24 horas	615	154,0	199,7	<0,003
Total	3745	128,3	141,9	<0,003

Tabela 2 - Aderência de cada período à distribuição exponencial (1h)

Período	Quantidade de chamados	Média de tempo entre chamados (minutos)	Desvio Padrão (minutos)	Aderência a Distribuição Exponencial (p-valor)
0-1 hora	127	244,2	229,9	<0,003
1-2 horas	105	217,3	169,2	0,013
2-3 horas	71	191,5	163,8	0,417
3-4 horas	81	194,0	151,6	0,011
4-5 horas	77	163,7	145,2	<0,003
5-6 horas	58	156,9	116,4	0,020
6-7 horas	82	140,2	109,3	0,080
7-8 horas	45	110,6	109,5	<0,003
8-9 horas	182	97,4	95,7	0,038
9-10 horas	212	101,3	110,9	0,561
10-11 horas	227	88,1	97,7	0,501
11-11 horas	226	108,1	102,5	0,884
12-13 horas	227	87,7	87,7	0,124
13-14 horas	210	98,5	97,2	0,226
14-15 horas	243	103,1	114,3	<0,003
15-16 horas	233	90,7	107,9	<0,003
16-17 horas	219	113,6	124,3	0,017
17-18 horas	229	110,4	141,6	0,005
18-19 horas	185	95,7	116,6	<0,003
19-20 horas	240	113,3	141,0	<0,003
20-21 horas	214	125,3	179,6	<0,003
21-22 horas	240	154,1	197,0	<0,003
22-23 horas	174	176,8	213,0	<0,003
23-24 horas	137	241,3	223,6	0,005
Total	3745	128,3	141,9	<0,003

A Tabela 3 apresenta o teste de aderência realizado para o tempo de serviço de cada ambulância, além do tempo total. Percebe-se que para todas as ambulâncias, a taxa de serviço obteve aderência à distribuição exponencial, mostrando que com um nível de significância de 99%, as mesmas podem ser consideradas exponenciais.

Tabela 3 - Tempo médio de serviço total (minutos)

Ambulância	Amostra	Tempo de serviço	Desvio Padrão (minutos)	Aderência Exponencial (p-valor)
USA0009	412	118,31	77,67	<0.003
USB7555	656	72,76	36,85	<0.003
USB7556	492	79,98	50,86	<0.003
USB7557	1079	64,27	45,17	<0.003
Total	2639	77,75	54,13	<0.003

3.2 MODELAGEM

Nesta seção, foram propostos quatro modelos utilizados para representar o sistema do SAMU. Os modelos se utilizam de modelos clássicos de filas encontrados na literatura e de algumas aproximações com o objetivo de aproximar os modelos do sistema estudado. A modelagem foi realizada com o objetivo de obter indicadores de desempenho para o sistema, como tempo em fila para cada classe de chamados e taxa de utilização do sistema. Todos os modelos foram desenvolvidos por meio da linguagem R no *software* RStudio.

3.2.1 MODELO 1

O Modelo 1 modela de maneira separada os atendimentos de alta prioridade realizados pela USA e os atendimentos de baixa prioridade realizados pelas três USBs. Esse modelo utiliza modelos de filas M(t)/M/c, em que as taxas de chegada e atendimento são markovianas e a quantidade de servidores para atendimento é fixa, além de assumi-los como homogêneos. Já as taxas de chegada mudam de acordo com o período do dia em que se encontram, caracterizando sistemas dependentes do tempo. Assim, as taxas de chegada (λ_ae λ_b) são tratadas como um conjunto de taxas de chegada para cada período do dia $\lambda_a = (\lambda_{a1}, \lambda_{a2}, \lambda_{a3}, \dots, \lambda_{ai})$ e $\lambda_b = (\lambda_{b1}, \lambda_{b2}, \lambda_{b3}, \dots, \lambda_{bi})$, em que *i* representa a quantidade de períodos.

Utilizou-se o princípio da modelagem do SIPP, que considera o sistema M(t)/M/c e divide a análise do mesmo em períodos que são tratados como independentes entre si. Desta forma, o modelo analisa cada período *i* como um sistema M/M/c, onde cada taxa de chegada é considerada como estacionária.

A equação (1) está relacionada ao cálculo da taxa de utilização do sistema para os sistemas de alta e baixa prioridade. Desta maneira, obtém-se a taxa de utilização ρ_k, em função da taxa de chegada λ_k,

da quantidade de ambulâncias (m_k) e da taxa média de atendimento da classe em questão (μ_k), para cada classe ou prioridade *k*.

$$\rho_k = (\frac{\lambda_k}{m_k * \mu_k}) \quad (1)$$

A equação (2) está relacionada à probabilidade de o sistema se encontrar vazio, enquanto as equações (3) e (4) representam a probabilidade P_n de o sistema estar com n clientes. Percebe-se uma distinção no cálculo do P_n quando o sistema apresenta fila, ou seja, quando a quantidade de clientes n no sistema é maior que a quantidade m de servidores. Já a equação (5) apresenta o cálculo do tempo em fila, indicador importante para avaliar o desempenho do sistema.

$$P_0 = (\frac{1}{\sum_{n=0}^{m-1} \frac{(\rho*m)^n}{n!} + \frac{(\rho*m)}{1}(\frac{1}{1-\rho})}) \quad (2)$$

$$P_n = \frac{(\rho * m)^n}{n!} * P_0\,, \qquad n = 1,2, \dots, m-1 \quad (3)$$

$$P_n = \frac{(\rho^n * m^m)}{m!} * P_0\,, \qquad n = m, m+1, \dots \quad (4)$$

$$W_q = (\frac{(\rho * m)^n * \rho * \lambda}{m!\,(1-\rho)^2}) \quad (5)$$

Visto isso, para a modelagem do atendimento de chamados básicos utilizou-se o modelo M(t)/M/3, a taxa de atendimento foi calculada fazendo-se uma média das três ambulâncias do sistema e a taxa de chegada foi calculada somando os atendimentos das três USBs. Para a USA, utilizou-se o modelo M(t)/M/1 e as taxas de atendimento e chegada foram obtidas diretamente da USA em cada período.

A taxa de utilização total do sistema foi calculada segundo a equação (6), onde assumiu-se que a taxa de utilização total é obtida pela soma das taxas de utilização obtidas individualmente nos modelos para alta (ρ_a) e baixa prioridade (ρ_b).

$$\rho = \rho_a + \rho_b \quad (6)$$

3.2.2 MODELO 2

O modelo 2 admite que as chamadas de alta prioridade devem ser atendidas por uma USB caso a USA esteja ocupada. Nesse modelo, os chamados de alta prioridade que aguardariam em fila são transferidos para serem atendidos por USBs. Para tal, continua-se tratando os modelos de alta e baixa

prioridade separadamente, porém, no atendimento de chamadas de alta prioridade não é admitida fila de espera e, com isso, o modelo utilizado passa a ser um M(t)/M/1/1. As chamadas perdidas nesse modelo são repassadas para o sistema que atende chamadas de baixa prioridade. Essa perda pode ser calculada pela probabilidade complementar da Equação (1), porém, por se tratar de um sistema com apenas um servidor em atendimento, a equação de perda pode ser resumida como visto em (7). Dessa forma, esta perda é adicionada ao vetor de chegada do sistema de baixa prioridade, que continua sendo modelado como M(t)/M/3.

$$Perda = \frac{\lambda_a}{\mu_a} \tag{7}$$

onde λ_a e μ_a são as taxas de chegada e atendimento da classe de alta prioridade.

Esse modelo também utiliza o SIPP tratando os períodos como independentes entre si. Assim, utiliza o modelo M/M/1/1 para os chamados e alta prioridade e o modelo M/M/3 para os chamados de baixa prioridade e para as chamadas de alta prioridade não atendidas pelo modelo anterior.

Os indicadores de desempenho são calculados de forma análoga às equações apresentadas no modelo 1, havendo apenas a mudança na taxa de chegada ao considerar a perda. Desse modo, a taxa de utilização do sistema de alta prioridade que está sendo modelado como M/M/1/1 passe a ser calculada como na equação (8).

$$\rho_a = \frac{\lambda_a - Perda}{\mu_a} \tag{8}$$

3.2.3 MODELO 3

O modelo 3 considera as prioridades entre os chamados ao selecionar o próximo chamado a ser atendido. Esse modelo utiliza disciplina não preemptiva, na qual o atendimento de um chamado de menor prioridade não é interrompido ao chegar ao sistema um chamado de maior prioridade.

Para tal, foi utilizado o modelo M(t)/M/c/NRRP e, em cada período, utilizando os princípios do SIPP o modelo M/M/c/NRRP (ARENALES et al., 2007). Esse modelo considera a prioridade, mas apresenta algumas limitações que o distanciam do sistema real. Primeiramente, não há como diferenciar os servidores, ou seja, as ambulâncias são assumidas como homogêneas, utilizando-se da taxa média de serviço de todas as ambulâncias. Dito isso, pode-se perceber que este modelo ilustra uma situação hipotética para esse sistema do SAMU na qual todas as ambulâncias são USAs.

As equações (6) a (9) apresentam o cálculo dos indicadores de desempenho para o modelo 3. A principal diferença em relação aos modelos anteriores está no tratamento dos servidores, onde é assumida uma taxa média de atendimento μ e uma quantidade de servidores m única para todo o sistema.

$$\rho_{k=}\left(\frac{\lambda_k}{m*\mu}\right) \tag{9}$$

$$\rho = \sum_{k=1}^{n} \rho_k \tag{10}$$

$$P_0 = \left(\frac{1}{\sum_{n=0}^{m-1}\frac{(\rho*m)^n}{n!} + \frac{(\rho*m)^m}{m}\left(\frac{1}{1-\rho}\right)}\right) \tag{11}$$

$$W_{qk} = \left(\frac{(\rho_k * m)^m * P_0}{m\mu(1-\rho_k)m!}\right) * \left(\frac{1}{(1-\rho)^2}\right) \tag{12}$$

3.2.4 MODELO 4

O modelo 4 trata as chamadas de alta prioridade independente das chamadas de baixa prioridade assim como no modelo 2. Entretanto, admite prioridade no atendimento das chamadas de alta prioridade quando atendidas por USBs assim como no Modelo 3. Assim, também utilizando os princípios do SIPP, para as chamadas de alta prioridade é utilizado o modelo M(t)/M/1/1 em cada período e a perda é calculada pela Equação (7). Esta perda é transferida para ser atendida juntamente com as chamadas de baixa prioridade, onde é utilizado o modelo M(t)/M/c/NRRP, em que há prioridade no atendimento das ocorrências, e o modelo M/M/c/NRRP em cada período.

Desta forma, os indicadores de desempenho do sistema relacionados ao modelo são calculados de forma análoga ao modelo 3, utilizando-se das equações (9) a (12) para o M/M/C/NRRP. Porém, o vetor de chegada da classe de alta prioridade é substituído pela perda encontrada na equação (7). Já para o modelo M/M/1/1, a taxa de utilização é calculada como em (8). Neste caso, também é feita a suposição de complementaridade das taxas de utilização, de modo que a taxa total é calculada pela soma da baixa com alta prioridade, como visto em (6).

3.3 RESULTADOS

A partir dos modelos construídos na seção anterior, procurou-se utilizar os resultados para comparar os modelos e o sistema em estudo. A análise de resultados da modelagem permitiu a comparação dos indicadores de desempenho do sistema com os dados obtidos diretamente. Desta forma, esta seção apresenta a validação dos modelos seguida de uma análise de sensibilidade dos parâmetros do sistema.

3.3.1 DIVISÃO DO DIA EM PERÍODOS DE 4 HORAS

Esta seção visa apresentar e analisar os resultados obtidos pela modelagem por meio do SIPP e SIPP PRI utilizando-se do período t equivalente a 4 horas, ou seja, o dia dividido em 6 períodos, onde considera-se a taxa de chegada como estacionária e distinta em cada um deles. Além disso, a seção propõe uma variante para o SIPP PRI baseada na revisão de literatura e compara seus resultados com a modelagem inicial.

3.3.1.1 TAXA DE UTILIZAÇÃO

Um fator muito importante para a análise de desempenho de um sistema de filas é a taxa de utilização dos servidores e do sistema como um todo. Os modelos propostos na seção anterior permitem o cálculo desta taxa como pode ser visto em (1), (6), (8), (9) e (10). Além disso, pode-se obter a taxa de utilização amostral do sistema realizando um cálculo simples de uma razão entre o tempo em que as ambulâncias estiveram ocupadas e o tempo total da análise.

A Tabela 4 apresenta um comparativo entre a taxa total de utilização do sistema obtida pelos modelos 1, 2, 3 e 4 e a taxa amostral. Os resultados sugerem que os modelos 1 e 4 são os que mais se aproximam da taxa amostral, porém apresentam desvio médio significativo de 29,6% e 52,7% em relação à amostra, respectivamente.

Tabela 4 - Taxa de utilização do sistema para cada modelo

Período	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Taxa Amostral
0-4 horas	0,1529	0,1057	0,0871	0,1337	0,1953
4-8 horas	0,1469	0,1002	0,0821	0,1287	0,2047
8-12 horas	0,3570	0,2336	0,1924	0,3015	0,4746
12-16 horas	0,4072	0,2543	0,2073	0,3383	0,5488
16-20 horas	0,4102	0,2431	0,1983	0,3379	0,4843
20-24 horas	0,3474	0,2127	0,1739	0,2922	0,4343

A discrepância observada entre a modelagem e os resultados obtidos através dos dados reais sugerem que a divisão do tempo em períodos pode estar causando uma perda dos chamados dentro dos subperíodos estudados. Com base nesta suposição, procurou-se aplicar uma variável do SIPP PRI que se ajustasse ao caso real e, consequentemente, à taxa amostral encontrada. A variante proposta foi baseada na LAG AVG SIPP PRI, vista em Vile et al. (2016) e explicada neste trabalho na seção 2.3.1. Por se tratar de uma variante aplicável à abordagem SIPP PRI, análoga ao modelo 4, a mesma foi aplicada apenas a este modelo, desconsiderando os demais por hora.

A Tabela 5 mostra o mesmo comparativo de taxa de utilização da Tabela 4 adicionando os resultados oriundos da aplicação da variante LAG AVG no modelo 4. Observa-se que o desvio médio de 52,7% entre o modelo e a taxa amostral, reduziu-se para 8,2% ao se fazer o comparativo dos novos resultados, o que sugere que esta abordagem tem impacto positivo significativo na modelagem e confirma a suposição de que a perda de chamados em função da divisão de períodos era o que estava influenciando na discrepância em relação à taxa de utilização amostral.

Tabela 5 - Taxa de utilização adicionando a variante LAG AVG

Período	Modelo 4	Modelo 4 com variante LAG AVG	Taxa amostral
0-4 horas	0,1337	0,1901	0,1953
4-8 horas	0,1287	0,2598	0,2047
8-12 horas	0,3015	0,4599	0,4746
12-16 horas	0,3383	0,5041	0,5488
16-20 horas	0,3379	0,4912	0,4843
20-24 horas	0,2922	0,3833	0,4343

A Figura 4 ilustra a visível aproximação entre a taxa calculada pela modelagem utilizando a variante (Modelo 4 com LAG AVG) em relação a amostra, além dos resultados para o Modelo 4. Esses resultados incentivam o uso da variante no modelo 4.

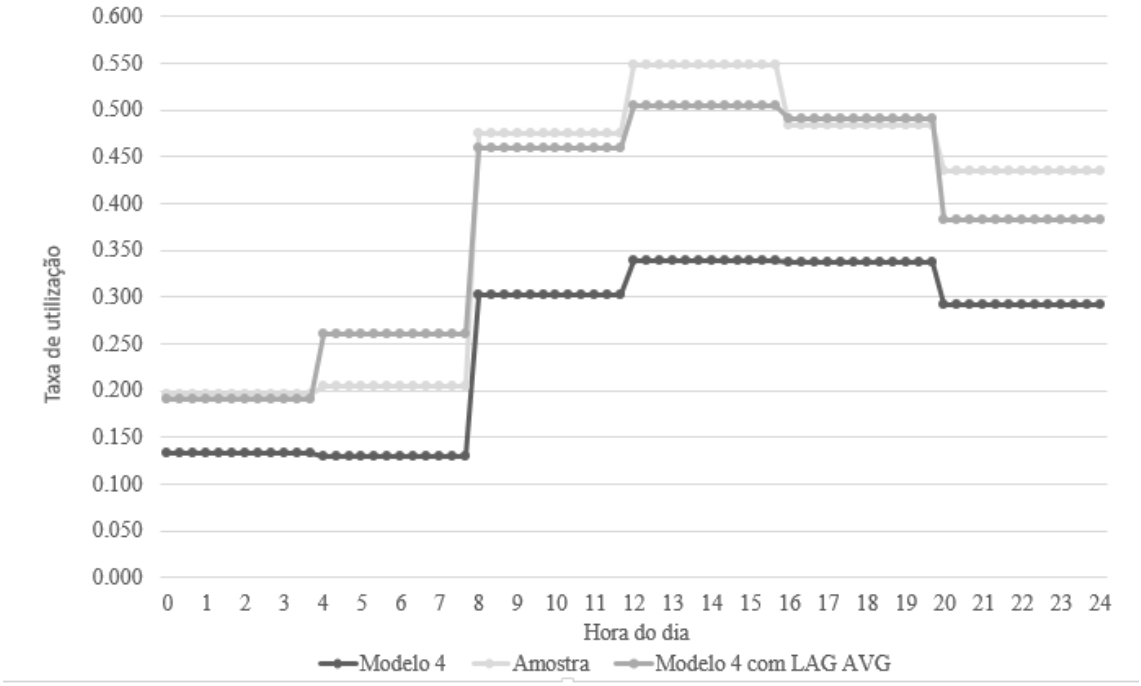

Figura 4 - Taxa de utilização com *t=4* versus hora do dia

3.3.1.2 TEMPO EM FILA

O tempo de resposta dos chamados é outro importante indicador de desempenho de sistemas emergências. Visto que utilizando esses modelos de filas não é possível calcular diretamente esse tempo, analisou-se o tempo em fila, já que o tempo de resposta nada mais é que o tempo que o servidor espera em fila adicionado ao tempo de deslocamento do mesmo. Nesse caso, pode-se inferir que o tempo de deslocamento dos servidores não se altera para todos os modelos, portanto o único fator que influencia no tempo de resposta a título de comparação entre os quatro tipos de modelagem é o tempo médio em fila.

A Tabela 6 apresenta o tempo médios que um paciente espera em fila para cada tipo de modelagem, período do dia e classe de atendimento. Pode-se perceber que, no geral, os dois primeiros períodos do dia, apresentam tempos em fila baixos e os três períodos seguintes apresentam valores mais altos, sugerindo que o sistema está mais congestionado.

Tabela 6 – Tempo em fila para cada modelo em minutos

Período	Tempo médio em fila para alta prioridade				Tempo médio em fila para baixa prioridade			
	Modelo 1	Modelo 2*	Modelo 3	Modelo 4**	Modelo 1	Modelo 2	Modelo 3	Modelo 4
0-4 horas	7,4748	-	0,0096	0,0738	0,0816	0,0894	0,0102	0,0816
4-8 horas	7,4748	-	0,0078	0,0612	0,0672	0,0738	0,0084	0,0672
8-12 horas	21,7542	-	0,1794	0,6588	0,7596	0,8472	0,2222	0,8256
12-16 horas	28,8114	-	0,2382	0,7656	0,8724	0,9948	0,3006	0,9708
16-20 horas	32,0952	-	0,2022	0,6222	0,7092	0,8268	0,2526	0,7758
20-24 horas	24,4944	-	0,1242	0,4470	0,5124	0,5880	0,1506	0,5432

*Classe não aceita filas no modelo.

**Resultados referentes aos chamados perdidos adicionados no modelo.

Em relação ao tempo em fila dos chamados de alta prioridade, os resultados apresentaram diferenças significativas entre os modelos. O modelo 1 apresenta resultados muito superiores aos demais modelos. Esse modelo utiliza apenas uma USA a qual apresenta tempo de atendimento alto. O modelo 2 não admite fila de espera e os chamados não atendidos são encaminhados para as USBs, porém esse atendimento é realizado sem diferenciar as classes dos chamados. O modelo 3 assume que existem quatro servidores homogêneos, ou seja, todos os servidores podem atender chamados dessa classe, além de prioridade entre as classes. O modelo 4 teoricamente se aproxima mais da realidade do sistema em estudo. Vale destacar que é preciso refinar os modelos para calcular o tempo em fila de chamados de alta prioridade.

Já para os chamados de baixa prioridade, o modelo 4 mostra tempos de fila próximos aos modelos 1 e 2, porém ligeiramente maiores em todos os períodos. Pelos mesmos motivos apresentados para os chamados de alta prioridade, os resultados do Modelo 3 se distanciam dos demais modelos.

Ao comparar-se os resultados de tempo em fila com os tempos de resposta do sistema, percebeu-se que, em geral, nos chamados de alta prioridade, o tempo em fila representa uma menor parcela do tempo de resposta, variando entre 0,37% em um período mais ocioso e 4,13% no período mais congestionado. Já na classe de baixa prioridade, esta mesma parcela variou de 0,42% até 6,82%. Os resultados sugerem que a fila tem mais impacto nos chamados de baixa prioridade, principalmente em períodos de pico.

Porém, como foi visto na seção anterior, os modelos de 1 a 4 apresentam uma taxa de utilização inferior àquela observada na amostra, o que pode estar causando uma situação análoga quanto aos

tempos em fila. Observou-se, também, que a abordagem com o uso da variante LAG AVG fez com que os resultados da modelagem para taxa de utilização se aproximassem muito mais da taxa encontrada de forma amostral, o que sugere que esta abordagem seja mais condizente com a realidade do sistema. Visto isso, optou-se por realizar uma análise dos tempos em fila para cada período do dia após a adição da variante no modelo, a fim de observar os novos resultados.

A Tabela 7 apresenta os tempos em fila já conhecidos para o modelo 4, além de adicionar os resultados da modelagem com a variante LAG AVG. Percebe-se que os novos tempos em fila, de maneira geral, apresentaram um acréscimo considerável. Para a fila de alta prioridade, observou-se que, em média, os tempos em fila utilizando a variante foram 2,7 vezes maiores que os tempos originais do modelo 4. Já para a fila de baixa prioridade, os tempos apresentaram valores em média 2,9 vezes maiores que os originais, sendo que, em alguns períodos com maior taxa de ocupação, estes valores foram maiores que 4 vezes os obtidos a partir do modelo 4.

Os resultados sugerem que, a partir do aumento na taxa de utilização do sistema com o emprego da variante LAG AVG, os tempos em fila obtiveram um aumento significativo, de forma que os valores para alguns períodos chegaram próximos ou até foram superiores a 3 minutos. Antes de realizar a modelagem com a variante, observou-se que os tempos em fila para os chamados representavam no máximo 6,82% dos tempos de resposta obtidos de forma amostral no sistema. Já após a inserção da variante, essa parcela do tempo em fila sobre tempo de resposta aumentou para até 18,7%.

Tabela 7 – Tempo em fila adicionando a variante LAG AVG em minutos

Período	Tempo médio em fila para alta prioridade		Tempo médio em fila para baixa prioridade	
	Modelo 4	Modelo 4 com variante LAG AVG	Modelo 4	Modelo 4 com variante LAG AVG
0-4 horas	0,0738	0,1274	0,0816	0,1437
4-8 horas	0,0612	0,2557	0,0672	0,2990
8-12 horas	0,6588	1,8255	0,8256	2,5238
12-16 horas	0,7656	2,1953	0,9708	3,0940
16-20 horas	0,6222	1,7925	0,7758	2,4710
20-24 horas	0,4470	0,8623	0,5432	1,1057

Dessa forma, percebe-se que o tempo em fila não é de fato um fator desprezível para a análise de desempenho do processo de atendimento emergencial do SAMU e que uma possível melhora nos resultados deste indicador para os períodos mais congestionados poderia ter impacto significativo nos tempos de resposta do sistema.

3.3.2 DIVISÃO DO DIA EM INTERVALOS DE 1 HORA

Visto que os resultados para a divisão do dia em 4 horas já foram apresentados, esta seção propõe-se a analisar e comparar os mesmos indicadores utilizando-se da divisão em 24 períodos de 1 hora cada, mantendo-se as mesmas características de modelagem. Neste caso, optou-se por analisar apenas os resultados para o modelo 4 e a variante associada devido à validação da mesma na seção anterior. Essa seção foi motivada por indicação da literatura (Vile et al., 2016) que sugere que diminuir os intervalos melhora a aderência dos modelos.

A taxa de utilização ou de ocupação do sistema pode ser calculada da mesma maneira sugerida na seção 5.1.1 para os períodos do dia divididos em intervalos de 1 hora. Como pode ser observado na Tabela 8, os resultados deste indicador utilizando o modelo 4 e o modelo 4 com a variante foram significativamente inferiores da taxa amostral ao considerar intervalos de 1 hora. Percebe-se que, para a modelagem inicial pelo modelo 4, os tempos encontrados foram em média 5,85 vezes menores em relação aos tempos da amostra, o que sugere que a divisão do dia em períodos menores de tempo distancia o modelo da realidade. Estes resultados obtidos sugerem a adequação da hipótese utilizada na seção 5.1.1 para justificar o emprego da variante LAG AVG SIPP PRI.

Contudo, quando a variante é aplicada para esta divisão de períodos, não acontece a mesma aproximação vista na divisão de $t = 4$, de modo que as taxas de ocupação continuam sendo 2,14 vezes menores do que às obtidas de forma amostral. Porém, percebe-se, novamente, uma melhora significativa dos resultados aplicando-se a variante, o que sugere que o principal problema seja realmente a divisão em períodos muito pequenos, apesar da abordagem da variante não ser capaz de solucioná-lo.

Tabela 8– Taxa de utilização para *t* = 1 hora

Período	Modelo 4	Modelo 4 com variante LAG AVG	Taxa amostral
0-1 hora	0,0410	0,1050	0,2578
1-2 horas	0,0393	0,0897	0,1937
2-3 horas	0,0272	0,0825	0,1536
3-4 horas	0,0284	0,0742	0,1765
4-5 horas	0,0308	0,0666	0.2244
5-6 horas	0,0199	0,0821	0,1149
6-7 horas	0,0242	0,1258	0,1746
7-8 horas	0,0559	0,1816	0,3051
8-9 horas	0,0674	0,2016	0,4333
9-10 horas	0,0805	0,2296	0,4857
10-11 horas	0,0787	0,2272	0,5250
11-11 horas	0,0898	0,2374	0,4543
12-13 horas	0,0839	0,2429	0,5179
13-14 horas	0,0841	0,2476	0,5223
14-15 horas	0,1007	0,2615	0,5087
15-16 horas	0,0934	0,2578	0,6463
16-17 horas	0,0982	0,2430	0,5635
17-18 horas	0,0925	0,2403	0,4450
18-19 horas	0,0764	0,2421	0,3923
19-20 horas	0,0990	0,2532	0,5366
20-21 horas	0,0931	0,2407	0,5125
21-22 horas	0,0920	0,2069	0,5838
22-23 horas	0,0708	0,1545	0,3542
23-24 horas	0,0543	0,1281	0,2868

A Figura 5 ilustra esta situação. Pode-se observar que a abordagem com variante aproxima a taxa de utilização obtida da taxa obtida diretamente dos dados, porém ainda com significativa discrepância entre os resultados, principalmente quando o sistema está mais congestionado.

Na Figura 6, pode-se observar o comparativo de como as curvas de taxa de utilização se comportam ao utilizar intervalos de 1 e 4 horas ao se adicionar a variante proposta à modelagem. A aderência à curva de taxa amostral ocorre apenas para a modelagem do período de 4 horas utilizando-se da variante.

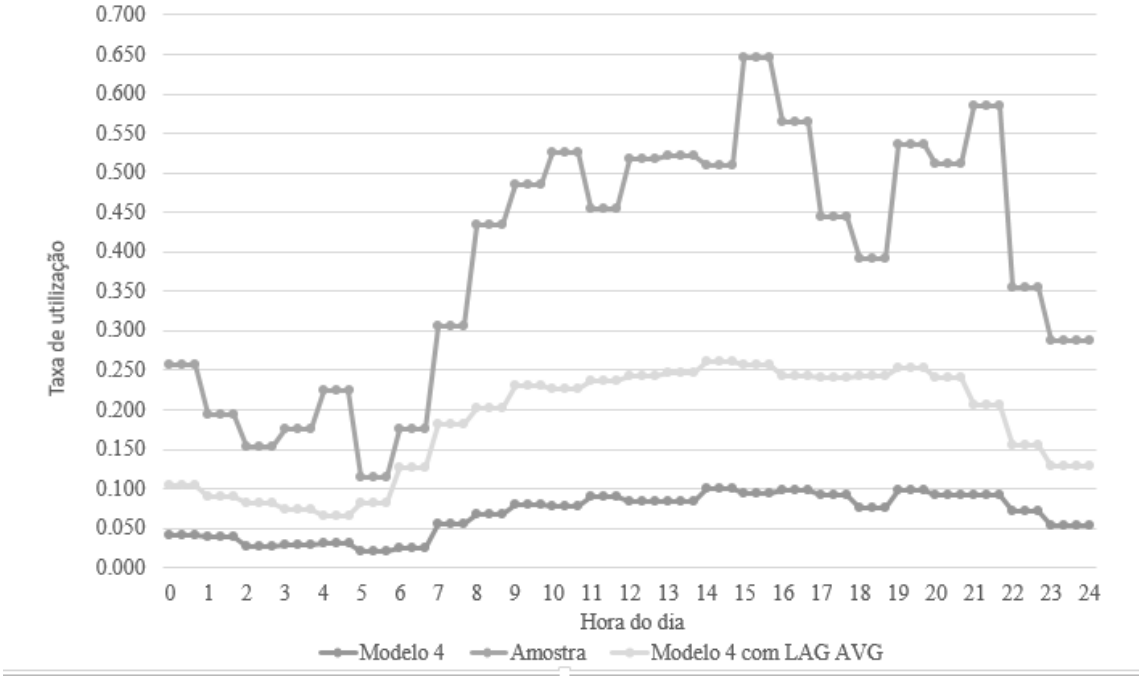

Figura 5 - Taxa de utilização com *t=1* versus hora do dia

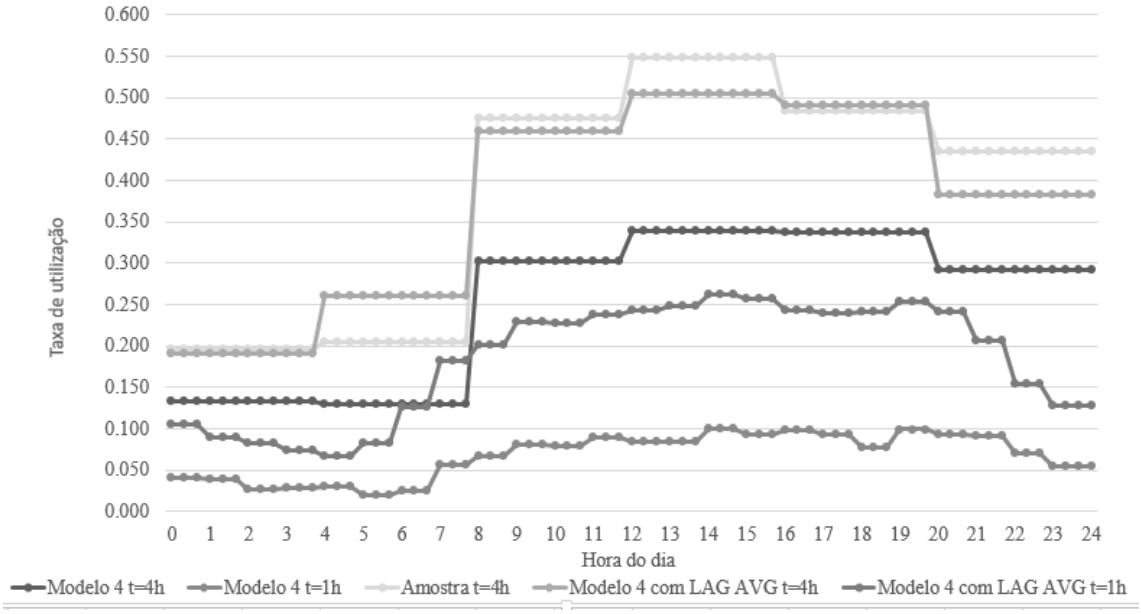

Figura 6 - Comparativo de taxas de utilização com *t*=1 e *t*=4 horas

3.3.3 ANÁLISE DE SENSIBILIDADE

Os resultados vistos na seção anterior sugerem que a divisão do dia em intervalos menores faz com que o modelo se distancie das características do sistema. Acredita-se que isto se deve ao fato de o tempo de atendimento ser consideravelmente grande. O tempo médio de atendimento das

ambulâncias está na ordem de 77 minutos, portanto ao dividir o dia em períodos de uma hora, o tempo de atendimento é maior que o intervalo utilizado na modelagem.

Esta seção se propõe a realizar uma análise de sensibilidade, variando o tempo de atendimento do sistema com o intuito de observar como os modelos se comportam, principalmente no que diz respeito à decisão de como dividir o dia em períodos menores.

A Figura 7 apresenta um comparativo semelhante aos vistos nas seções anteriores para a taxa de ocupação do sistema, porém, neste caso, o tempo de atendimento de todas as ocorrências do sistema foram divididas por 4, com o intuito de simular a aderência da modelagem a um sistema que apresente uma taxa de atendimento 4 vezes maior. É possível observar que a mudança no tempo de atendimento resulta em uma melhora dos resultados em relação à taxa amostral, porém não sugere que a divisão do dia em períodos de 1 hora seja mais adequada neste caso. Na realidade, a mudança na taxa de atendimento resultou em uma melhoria média de 5,1% nos resultados para t=4h, enquanto a melhoria para t=1h foi de apenas 1,2% em média.

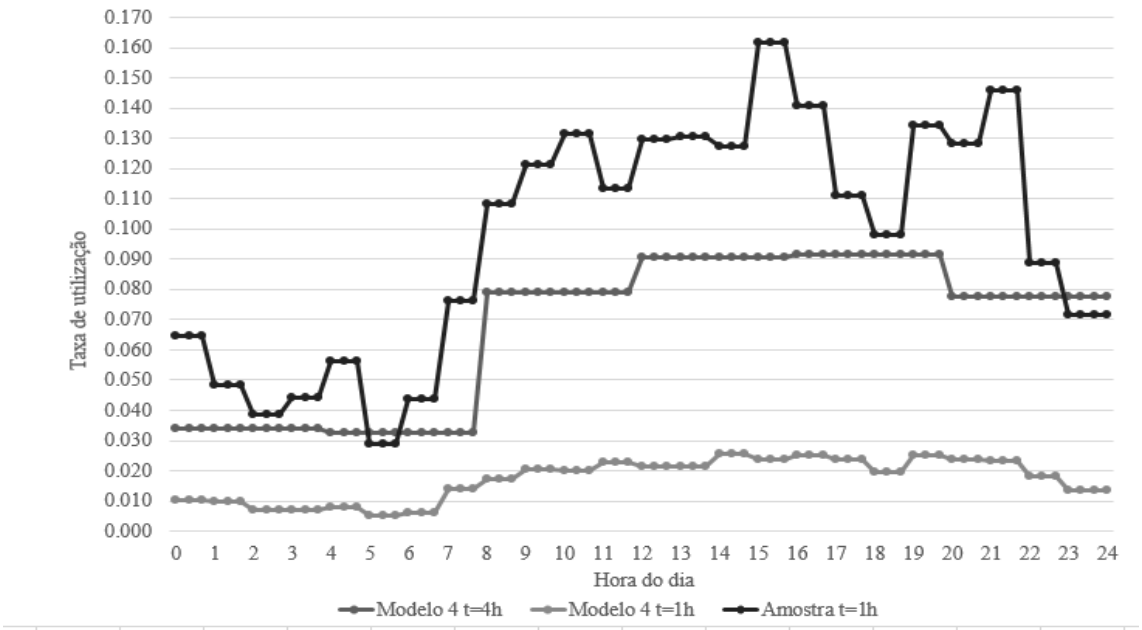

Figura 7 - Comparativo de taxas de utilização com *t*=1 e *t*=4 horas

A mesma situação ocorre para o caso da modelagem utilizando-se da variante proposta com a nova taxa de atendimento, de modo que não há como inferir que a divisão do dia em períodos menores possa ajudar na aderência da modelagem em relação ao sistema real. Conclui-se que, para os moldes propostos, a modelagem que mais se adequa ao sistema real é a do modelo 4 com variante LAG AVG

e utilizando-se de um *t*=4h, portanto, os tempos em fila a serem utilizados como indicadores de desempenho do sistema SAMU devem ser os vistos na Tabela 7.

3.4. ANÁLISE ECONÔMICA

Esta seção se propõe a analisar os cenários observados nas modelagens de acordo com a perspectiva econômica, comparando os resultados dos indicadores de desempenho com os custos atrelados de se manter um sistema com determinadas características.

3.4.1 FUNÇÃO CUSTO

Uma função de custo foi elaborada com o objetivo de descrever matematicamente os custos envolvidos na aquisição e manutenção de ambulâncias no sistema SAMU. A função geral abrange apenas a compra e manutenção da unidade ambulatória, além dos salários das equipes diretamente envolvidas com o atendimento de ocorrências, de forma que foram desconsiderados os demais custos operacionais do sistema, tais como regulação dos atendimentos e administração do sistema.

Como mencionado anteriormente, existem dois tipos de unidades ambulatórias no sistema SAMU. Essas unidades se distinguem pela gama de equipamentos presentes nas mesmas, além de uma diferenciação na equipe responsável. Segundo a Portaria Nº1010 de 21 de Maio de 2012 do Ministério da Saúde, são elas:

> I - Unidade de Suporte Básico: tripulada por no mínimo 2 (dois) profissionais, sendo um condutor de veículo de urgência e um técnico ou auxiliar de enfermagem;
>
> II - Unidade de Suporte Avançado: tripulada por no mínimo 3 (três) profissionais, sendo um condutor de veículo de urgência, um enfermeiro e um médico.

Percebe-se, portanto que os tipos de ambulâncias diferem tanto na unidade em si quanto nas equipes que atuam em seu serviço, portanto apresentam custos distintos de aquisição e manutenção no sistema, o que evidencia que as mesmas precisam ser tratadas com funções de custos distintas. Desta forma, a função custo foi elaborada em função do tipo de unidade ambulatória correspondente, sendo que os custos totais do sistema são obtidos pelo somatório entre a quantidade de ambulâncias para cada tipo.

A Equação (13) apresenta a função geral de custo em função dos custos individuais de quantidade de unidades ambulatórias, enquanto as equações (14) e (15) introduzem as funções relacionadas a cada

tipo de ambulância do sistema, onde o conjunto n está relacionado às Unidades de Serviço Avançado (USAs) e o conjunto m às Unidades de Serviço Básico (USBs). Na equação (14), o custo de manutenção A_i, para cada ambulância i, foi dividido por 12 meses uma vez que este custo está relacionado ao período de 1 ano, ao passo que o custo de aquisição da ambulância B_i é dividido por 12 meses vezes t, que representa a vida útil da ambulância. O valor salarial do condutor (C) é multiplicado pelo fator 4, pelo fato destes profissionais trabalharem em turnos de 12 horas de trabalho e 36 de descanso, de modo que são necessários 4 trabalhadores para manter uma ambulância em funcionamento 24 horas por dia. Já os valores de salário de enfermeiros e médicos estão multiplicados pelo fator 7, sendo que os mesmos atuam com uma carga horária de 24 horas semanais. No caso da equação (15), a única diferença acontece para os técnicos de enfermagem, que, assim como os motoristas, trabalham em turnos de 12 por 36 horas, recebendo um fator multiplicados de 4 vezes.

$$Z = \sum_{i=1}^{n} Q_i + \sum_{j=1}^{m} Q_j \tag{13}$$

$$Q_i = \left(\frac{A_i}{12}\right) + \left(\frac{B_i}{t * 12}\right) + 4 * C + 7 * (D + E) + F(\alpha) \tag{14}$$

$$Q_j = \left(\frac{A_j}{12}\right) + \left(\frac{B_j}{t * 12}\right) + 4 * (C + G) + F(\alpha) \tag{15}$$

Onde:

Z = Custo total mensal das ambulâncias do sistema

n = Número de Unidades de Serviço Avançado (USA)

m =Número de Unidades de Serviço Básico (USB)

Q_i = Custo mensal de uma Unidades de Serviço Avançado (USA)

Q_j = Custo mensal de uma Unidades de Serviço Básico (USB)

A_i = Custo de manutenção anual de uma Unidades de Serviço Avançado (USA)

A_j = Custo de manutenção anual de uma Unidades de Serviço Básico (USB)

B_i = Custo de aquisição anual de uma Unidades de Serviço Avançado (USA)

B_j = Custo de aquisição anual de uma Unidades de Serviço Básico (USB)

C = Custo com salário do motorista

D = Custo com salário do enfermeiro

E = Custo com salário do médico

G = Custo com salário do técnico de enfermagem;

F(α) = Função de custo com combustível de acordo com quilometragem rodada

Em seguida, partiu-se para a obtenção dos valores de custos utilizados na função. Os dados estão especificados na Tabela 9, assim como as respectivas fontes. Os dados de salários foram obtidos a partir de um edital de contratação do consórcio Aliança, responsável pela regulação do sistema SAMU Ouro Preto, os valores de aquisição de unidade ambulatória foram retirados do site do governo, que realiza venda das mesmas, o custo de manutenção foi retirado de uma licitação de um município do estado do Paraná, além do preço do diesel, que foi obtido em um site de comparação de preços de combustíveis ao longo do tempo. É importante ressaltar que os valores são aproximações e não representam com exatidão os custos do sistema, porém considerou-se a melhor opção de estimativa devido à dificuldade de obtenção direta destes dados com a gerência do sistema.

Tabela 9 – Valores dos custos relacionados ao SAMU e fontes

Tipo de Custo	Valor	Fonte
Manutenção Ambulância	R$ 8400,00 (anual)	http://www.chopinzinho.pr.gov.br/portal/licitacoes/1458669307.pdf
Aquisição USA	R$ 290.000,00	https://portalgoverno.com.br/product/ambulancia-de-suporte-avancado-tipo-d/
Aquisição USB	R$ 178.000,00	https://portalgoverno.com.br/product/ambulancia-de-resgate-tipo-c/
Salário Motorista	R$ 1.497,29	https://www.msconcursos.com.br/concurso/664001/consrcio-intermunicipal-aliana-para-a-sade-cias-mg
Salário Enfermeiro	R$ 2.101,07	https://www.msconcursos.com.br/concurso/664001/consrcio-intermunicipal-aliana-para-a-sade-cias-mg
Salário Técnico de Enfermagem	R$1.151,67	https://www.msconcursos.com.br/concurso/664001/consrcio-intermunicipal-aliana-para-a-sade-cias-mg
Salário Médico	R$ 6.287,37	https://www.msconcursos.com.br/concurso/664001/consrcio-intermunicipal-aliana-para-a-sade-cias-mg
Diesel	R$ 4,05 (litro)	https://precodoscombustiveis.com.br/pt-br/city/brasil/minas-gerais/ouro-preto/2782

De acordo com as equações (13), (14) e (15), percebe-se que todos os custos relacionados ao custo total de manter uma ambulância atuando no sistema são fixos, ou seja, não mudam de acordo com a base em que a ambulância está localizada nem com a sua utilização, com exceção da função de combustível F(α) que retorna um valor de custo com combustível de acordo com a quantidade de quilômetros que a ambulância percorre. Nessa função, foram utilizados o tempo de deslocamento de cada ambulância nessa análise. Foi realizado, portanto, um cálculo dos tempos totais de deslocamento de cada ambulância por mês do ano, como pode ser observado na Tabela 10.

Tabela 10 – Tempo de deslocamento mensal por ambulância em minutos

Mês/Ambulância	USA 0009	USB 7555	USB 7556	USB 7557
Janeiro	1946,499	6340,024	5561,424	6451,252
Fevereiro	2447,027	4727,211	4615,982	5895,110
Março	3114,398	4727,211	4171,068	6006,338
Abril	2891,941	4115,454	4226,682	7063,009
Maio	3114,398	4393,525	4282,297	6840,552
Junho	2725,098	4615,982	4727,211	7674,766
Julho	3559,312	5338,967	5338,967	7007,395
Agosto	3170,012	4727,211	4949,668	5728,267
Setembro	4004,226	6117,567	4782,825	6562,481
Outubro	3670,540	6451,252	4004,226	7007,395
Novembro	3837,383	6173,181	4171,068	6784,938
Dezembro	3281,240	6729,323	3837,383	6173,181

Como apenas encontravam-se disponíveis os dados de deslocamento das ambulâncias para o período de 12 meses, considerada uma amostra pequena para inferir o comportamento das mesmas, optou-se por realizar a análise de custo utilizando o método de Monte Carlo. Os tempos de deslocamento de cada ambulância foram assumidos como distribuídos uniformemente entre seus valores máximos e mínimos mensais encontrados. Para a USA 0009, por exemplo, esses valores estavam entre 1946,499 e 4004,226 minutos.

O método utilizado foi baseado em Andrade (2009) e consiste em utilizar uma função cumulativa uniforme para cada uma das ambulâncias. Em seguida, são gerados 10.000 números aleatórios, relacionando-os à variáveis de decisão, no caso o tempo de deslocamento. Desta forma, pode-se realizar uma estimativa mais precisa do tempo em que a ambulância se encontra em deslocamento no período de 1 mês.

Para relacionar os dados obtidos com o custo da ambulância com combustível, todos os valores foram divididos pela velocidade média de deslocamento e multiplicados pela média de consumo além do valor do combustível. Foram considerados para esta análise 60km/hora, 7km/litro e R$4,05/litro, respectivamente. Desta forma, obtém-se, portanto, o valor da função deslocamento para cada ambulância do sistema, que deve ser somado ao valor fixo associado ao tipo de unidade para se obter o custo total por ambulância. Estes valores podem ser observados na Tabela 11.

Tabela 11 – Custos Mensais de cada ambulância

Ambulância	Custo Fixo	Custo com Combustível	Custo Total
USA0009	R$70.241,57	R$1.722,94	R$71.964,52
USB7555	R$14.262,50	R$3.063,64	R$17.326,15
USB7556	R$14.262,50	R$2.950,58	R$17.213,09
USB7557	R$14.262,50	R$3.870,72	R$18.133,23
Total	R$113.029,09	R$11.607,90	R$124.637,00

3.4.2 ANÁLISE DOS CENÁRIOS

Esta seção tem como objetivo comparar os custos dos cenários e analisar o impacto ao mudar algumas características do sistema. Na Tabela 11, já é possível observar que o custo total médio do sistema com ambulâncias e equipes de atendimento é de R$124.637.00 mensais. Os cenários propostos são:

- Cenário 1: Equipar todas as ambulâncias, ou seja, transformar todas as USBs em USAs;
- Cenário 2: Diminuir uma ambulância USB do sistema nos períodos de ociosidade (0h às 8h);
- Cenário 3: Inserir uma ambulância USB no sistema nos períodos de congestionamento (8h às 24h);

A Tabela 12 mostra um comparativo dos cenários envolvendo as taxas de utilização do sistema e o custo mensal de cada cenário. As taxas de utilização foram calculadas baseando-se no modelo 4 com a variante proposta. Além disso, a última linha diz respeito ao custo total mensal de manter o sistema funcionando nos moldes estabelecidos para o cenário. Os valores de custo foram calculados de forma análoga à Tabela 11. A partir dos resultados, pode-se observar que o Cenário 1, onde todas as ambulâncias foram equipadas como USAs, apresentou as menores taxas de ocupação para todos os períodos do dia, porém o custo total do cenário mais que dobrou em relação ao inicial. No caso do cenário 2, a diminuição de uma ambulância nos dois primeiros períodos acarretou em um aumento médio de 39% na taxa de utilização destes períodos, porém mantendo-as em um nível abaixo dos

demais períodos. O custo, por sua vez, teve um decréscimo de apenas R$3.632,49 em relação ao cenário inicial. Finalmente, no cenário 3, o saldo foi de uma diminuição de 19,1% da taxa de utilização nos períodos de pico e um aumento de R$13.656,97 no custo mensal do sistema.

Tabela 12– Comparativo de cenários utilizando a taxa de utilização e custo mensal

Período	Cenário Inicial	Cenário 1	Cenário 2	Cenário 3
0-4 horas	0,1901	0,1072	0,2384	0,1901
4-8 horas	0,2598	0,1399	0,3216	0,2598
8-12 horas	0,4599	0,2609	0,4599	0,3915
12-16 horas	0,5041	0,2775	0,5041	0,4322
16-20 horas	0,4912	0,2648	0,4912	0,4232
20-24 horas	0,3833	0,2105	0,3833	0,3288
Custo Mensal	R$124.637,00	R$287.858.09	R$121.004,51	R$138.293,97

Além da taxa de utilização, o fator mais importante a ser comparado com os custos nos cenários propostos é o tempo médio em fila, indicador que tem impacto direto no tempo de resposta às ocorrências. A Tabela 13 apresenta os resultados deste indicador para os cenários propostos, comparando os tempos obtidos na modelagem do cenário inicial que simula as características reais do sistema. Os resultados do cenário 1 se mostraram inferiores ao inicial, porém seu custo se mostrou muito elevado. Ressalta-se os resultados obtidos para o cenário 3. A inserção de uma ambulância extra nos períodos de pico acarretou em uma diminuição de mais de 9 vezes do tempo em fila obtido para o cenário inicial, além de deixar todos os tempos em fila na mesma ordem de grandeza, abaixo do 0,33 minutos. No caso do cenário 2, também se observou um resultado parecido, no qual os tempos em fila se mostraram da mesma ordem de grandeza, porém com valores de até 3,59 minutos. Estes dois cenários representam dois casos possíveis de implementação no sistema ao considerar variação na quantidade de equipes atuando de acordo com os horários do dia. A decisão de qual dos dois cenários aplicar está associada com o nível de serviço aceitável para o SAMU.

Tabela 13 - Comparativo de Cenários com Tempo de Fila

Período	Tempo médio em fila para alta prioridade				Tempo médio em fila para baixa prioridade			
	Cenário Inicial	Cenário 1	Cenário 2	Cenário 3	Cenário Inicial	Cenário 1	Cenário 2	Cenário 3
0-4 horas	0,1274	0,0208	1,6782	0,1274	0,1437	0,0233	2,0021	0,1437
4-8 horas	0,2557	0,0559	2,8482	0,2557	0,2990	0,0650	3,5909	0,2990
8-12 horas	1,8255	0,5928	1,8255	0,2137	2,5238	0,8021	2,5238	0,2699
12-16 horas	2,1953	0,7837	2,1953	0,2582	3,0940	1,0847	3,0940	0,3305
16-20 horas	1,7925	0,6394	1,7925	0,2091	2,4710	0,8697	2,4710	0,2635
20-24 horas	0,8623	0,2529	0,8623	0,0920	1,1057	0.3203	1,1057	0,1103

4. CONCLUSÕES

Esse trabalho analisou o sistema do SAMU Ouro Preto e Mariana utilizando-se de conceitos de Teoria de Filas e da abordagem SIPP PRI, propondo modelos e variantes que permitem estudar sistemas de filas envolvendo prioridades e dependentes do tempo. A abordagem proposta aqui difere das demais encontradas na literatura por propor um modelo baseado nos conceitos do SIPP PRI que permite calcular diretamente indicadores de desempenho como o tempo médio em fila e a taxa de utilização, além de comparar os resultados da modelagem com os indicadores amostrais do próprio sistema.

Os resultados comprovaram que o sistema funciona de maneira homogênea em relação aos meses do ano e dias da semana, porém de maneira heterogênea durante os períodos do dia. A modelagem utilizando-se da abordagem SIPP permitiu mostrar as diferenças de ocupação dos recursos ao longo dos períodos do dia, além de mostrar os horários de pico e ociosidade do sistema, proporcionando uma ferramenta adequada para auxílio a tomada de decisão dos gestores do mesmo.

Além disso, foram propostos cenários alternativos para avaliar os impactos econômicos de mudanças de determinadas características do sistema utilizando o método de Monte Carlo. Nessa análise, foram utilizados os custos de manutenção dos veículos, equipes, aquisição da ambulância e combustível. Essas análises também permitiram comparar os custos relacionados aos cenários propostos com as mudanças nos indicadores de desempenho do sistema. Destaca-se o Cenário 3, que propõe a inclusão de mais uma ambulância nos horários de pico de atendimentos, o que resultaria em uma diminuição de mais de 9 vezes o tempo em fila em relação ao cenário inicial. Além disso, é possível utilizar o cenário 2, caso a gerencia do sistema opte por admitir um nível de serviço menor.

Para trabalhos futuros, recomenda-se a comparação dos resultados obtidos na modelagem com métodos exatos como EULER PRI e modelos de simulação, além da aplicação dos modelos em outros sistemas reais ou fictícios, realizando uma análise de sensibilidade mais robusta. Podem também ser propostos mais cenários para o sistema SAMU, a fim de observar como o mesmo se comporta em outras situações não descritas. Por fim, recomenda-se também o acompanhamento dos dados dos próximos anos, podendo-se realizar um comparativo de demanda e uma avaliação de tendência para o futuro, utilizando-se da modelagem proposta para simular situações previstas.

REFERÊNCIAS

ANDRADE, Eduardo Leopoldo de. Introdução à Pesquisa Operacional: métodos e modelos para análise de decisões. Rio de Janeiro, Editora LTC, 2009.

ARENALES, Marcos, ARMENTANO, Vinícius, MORABITO, Reinaldo, YANASSE, Horácio. Pesquisa Operacional, Rio de Janeiro, Editora Campus, 2007.

DECRETO Nº 5.055, DE 27 DE ABRIL DE 2004. Palácio do Planalto. Acesso em 19 de outubro de 2015. Disponível em: http://www.planalto.gov.br/ccivil_03/_Ato2004-2006/2004/Decreto/D5055.htm

GREEN, Linda V., KOLESAR, Peter J., SOARES, João. Improving the Sipp Approach for Staffing Service Systems That Have Cyclic Demands. Operations Research, p. 549-564, 2001. Disponível em: http://dx.doi.org/10.1287/opre.49.4.549.11228.

GREEN, Linda V. Queueing theory and modeling. Graduate School of Business, Columbia University, New York, New York, 16p., 2011.

GREEN, Linda V., KOLESAR, Peter J. The pointwise stationary approximation for queues with nonstationary arrivals. Management Science, v. 37, n. 1, 1991.

Instituto Brasileiro de Geografia e Estatística (IBGE). Disponível em: https://cidades.ibge.gov.br. Consultado em 10 de abril de 2019.

IANNONI, Ana Paula, MORABITO, Reinaldo. Modelo de fila hipercubo com múltiplo despacho e backup parcial para análise de sistemas de atendimento médico emergenciais em rodovias. Pesquisa Operacional, v. 26, n. 3, p. 493–519, 2006.

INGOLFSSON, Armann, AKHMETSHINA, Elvira, BUDGE, Susan, LI, Yongyue, WU, Xudong. A survey and experimental comparison of service-level-approximation methods for nonstationary M(t)/M/s(t) queueing systems with exhaustive discipline. INFORMS Journal on Computing, v. 19, n. 2, p. 201–214, 2007.

IZADY, Navid. On queues with time-varying demand, PhD thesis, Lancaster University Management School, 2010.

IZADY, Navid, WORTHINGTON, Dave. Setting staffing requirements for time dependent queueing networks: The case of Accident and Emergency departments. European Journal of Operational Research, p. 219, n. 3, p. 531–540, 2012.

Jornal Estado de Minas. Disponível em: https://www.em.com.br/app/noticia/gerais/2014/04/24/interna_gerais,522063/samu-registra-demora-media-de-atendimento-30-superior-ao-maximo-recomendado.shtml. Acesso em: 04 de janeiro de 2018.

MCLAY, Laura A., MAYORGA, Maria E. Evaluating emergency medical service performance measures, Health Care Management Science, v. 13, n. 2, p. 124-136, 2009.

SCHWARZ, Justus Arne, SELINKA, Gregor, STOLLTRZ, Raik. Performance analysis of time-dependent queueing systems: Survey and classification. Omega. p. 170-185, 2016.

SILVA, Glauco. Processos markovianos intermediários: um modelo de filas com numero variável de servidores. Dissertação - Universidade Federal do Rio de Janeiro, COPPE, 2008.

STOLLETZ, Raik. 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. p. 478–493, 2008.

STOLLETZ, Raik, Analysis of passenger queues at airport terminals. Research in Transportation Business & Management, v. 1, p. 144–149, 2011.

STOLLETZ, Raik, LAGERSHAUSEN, Svenja. Time-dependent performance evaluation for loss-waiting queues with arbitrary distributions. International Journal of Production Research. p. 1366-1378, 2013.

TAKEDA, Renata Algisi, WIDMER, João Alexandre, MORABITO, Reinaldo. Uma proposta alternativa para avaliação do desempenho de sistemas de transporte emergencial de saúde brasileiros. Transportes, v. 9, n. 2, p. 9-27, 2001.

VILE, Julie Leanne, GILLARD, Jonathan William, HARPER, Paul Robert, KNIGHT, Vincent Anthony. A queueing theoretic approach to limit excessive waiting times in time-dependent dual-class service systems, 2015, submetido a publicação. Disponível em: http://www.julievile.co.uk/publications.html.

VILE, Julie Leanne, GILLARD, Jonathan William, HARPER, Paul Robert, KNIGHT, Vincent Anthony. Time-dependent stochastic methods for managing and scheduling Emergency Medical Services. Operations Research for health care, v. 8, p. 42-52, 2016.

VILE, Julie Leanne, GILLARD, Jonathan William, HARPER, Paul Robert, KNIGHT, Vincent Anthony. A Queueing Theoretic Approach to Set Staffing Levels in Time-Dependent Dual-Class Service Systems. Decision Sciences, v. 48, n. 4, p. 766-794, 2017.

VUKMIR, Rade. Survival from prehospital cardiac arrest is critically dependent upon response time, Ressucitation, v. 69, n. 2, p. 269-334, 2006.