Ministério da Educação
Universidade Federal de Ouro Preto
Escola de Minas
Programa de Pós-Graduação em Engenharia de Produção

ANÁLISE DO SAMU DE OURO PRETO E MARIANA: UMA ABORDAGEM UTILIZANDO O MODELO HIPERCUBO DE FILAS INTEGRADO À META-HEURÍSTICA VNS

RAMON SÁVIO MOREIRA LOPES

Ouro Preto MG

2021

RAMON SÁVIO MOREIRA LOPES

ANÁLISE DO SAMU DE OURO PRETO E MARIANA: UMA ABORDAGEM UTILIZANDO O MODELO HIPERCUBO DE FILAS INTEGRADO À META-HEURÍSTICA VNS

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do título de Mestre em Engenharia de Produção.

Orientadora: Prof.ª Dr.ª Lásara Fabrícia Rodrigues
Coorientador: Prof. Dr. Helton Cristiano Gomes

Ouro Preto – MG
Agosto de 2021

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

L864a Lopes, Ramon Savio Moreira.
Análise do SAMU de Ouro Preto e Mariana [manuscrito]: uma abordagem utilizando o modelo hipercubo de filas integrado à meta-heurística VNS. / Ramon Savio Moreira Lopes. - 2021.
98 f.: il.: color., gráf., tab..

Orientadora: Profa. Dra. Lasara Fabricia Rodrigues.
Coorientador: Prof. Dr. Helton Cristiano Gomes.
Dissertação (Mestrado Acadêmico). Universidade Federal de Ouro Preto. Departamento de Engenharia de Produção. Programa de Pós-Graduação em Engenharia de Produção.

1. Hipercubo - Modelo de Filas. 2. Variable Neighborhood Search (VNS). 3. Servidores Emergenciais. 4. Serviço de Atendimento Móvel de Urgência (SAMU). I. Gomes, Helton Cristiano. II. Rodrigues, Lasara Fabricia. III. Universidade Federal de Ouro Preto. IV. Título.

CDU 658.5

Bibliotecário(a) Responsável: Maristela Sanches Lima Mesquita - CRB-1716

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
ESCOLA DE MINAS
DEPARTAMENTO DE ENGENHARIA DE PRODUCAO, ADMINISTRACAO E ECON

FOLHA DE APROVAÇÃO

Ramon Sávio Moreira Lopes

Análise do Samu de Ouro Preto e Mariana: uma abordagem utilizando o modelo Hipercubo de Filas integrado à meta-heurística VNS

Dissertação apresentada ao Programa de Pós-graduação em Engenharia de Produção (PPGEP) da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de mestre.

Aprovada em 29 de junho de 2021.

Membros da banca

Doutora - Lásara Fabrícia Rodrigues - Orientadora (Universidade Federal de Minas Gerais)
Doutor - Helton Cristiano Gomes - Coorientador (Universidade Federal de Ouro Preto)
Doutor - Reinaldo Morabito - (Universidade Federal de São Carlos)
Doutor - Frederico Rodrigues Borges da Cruz – (Universidade Federal de Minas Gerais)

Lásara Fabrícia Rodrigues, orientadora do trabalho, aprovou a versão final e autorizou seu depósito no Repositório Institucional da UFOP em 01/09/2021.

Documento assinado eletronicamente por **Helton Cristiano Gomes**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 01/09/2021, às 08:26, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0215593** e o código CRC **67649F06**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35400-000
Telefone: 3135591540 - www.ufop.br

“Mais importante do que a chegada é a trajetória
e quem acompanha a gente no caminho”
(Autor desconhecido)

Agradecimentos

A Deus, por me abençoar e iluminar durante essa trajetória.

Aos meus pais, Andréia e Savinho, por serem minhas principais fontes de inspiração e por me apoiarem em todos os momentos.

À minha irmã, Tamara, por ser a pessoa que eu mais convivi durante o período de elaboração dessa Dissertação, pela amizade e pelo incentivo.

À Profa. Lásara Rodrigues, por sempre acreditar em mim, pela amizade, orientação e dedicação.

Ao meu cunhado, Álisson, pelo convívio e por ter me auxiliado tanto ao longo da etapa de programação matemática.

Aos meus familiares, em especial à minha avó, Dona Imaculada, exemplo de resiliência e de determinação.

Ao Prof. Helton Gomes, pela coorientação, conselhos e amizade.

Aos demais professores do PPGEP-UFOP, por todos os ensinamentos e por serem exemplos de dedicação.

Aos colegas do PPGEP-UFOP, pela parceria e assistências nessa caminhada.

Aos professores Reinaldo Morabito e Frederico Rodrigues Borges da Cruz, pela disponibilidade, sugestões e críticas durante a Qualificação e Defesa da Dissertação, que colaboraram para aperfeiçoar esse trabalho.

Aos colegas e professores do PPGEP da UFMG, pelos ensinamentos e companheirismo durante o período das disciplinas isoladas, em especial à Virgínia, Guilherme Ferreira, Débora, Luiz, Higor, Luíza e Prof. Luiz Ricardo.

Aos meus amigos, em especial ao Rodrigo, Guilherme de Moura, André e Lucas, que contribuíram muito para o meu crescimento.

Aos colegas de trabalho da EY, pela parceria diária e por serem compreensivos durante a elaboração desse trabalho.

À FAPEMIG e à CAPES por acreditarem nesse trabalho e fomentarem esse sonho.

Resumo

Diariamente, indivíduos são submetidos a situações em que serviços básicos de saúde precisam ser requisitados. Um importante exemplo é o Serviço de Atendimento Móvel de Urgência (SAMU). Para que o SAMU mantenha ou melhore o nível do seu serviço, o tempo entre o recebimento da ocorrência e a chegada da unidade ao local da mesma deve ser o menor possível. Nesse trabalho foi desenvolvida uma ferramenta que incorporou o Modelo Hipercubo de Filas (MHF) à meta-heurística *Variable Neighborhood Search* (VNS), permitindo analisar as particularidades das políticas de despacho e localização dos servidores. O MHF trata do planejamento de sistemas de atendimento nos quais servidores se deslocam até clientes para prestar serviços e a escolha do VNS justifica-se por sua extensa utilização na resolução de diversos tipos de problema de localização. O Algoritmo VNS/MHF elaborado foi utilizado para otimizar a localização das ambulâncias do SAMU de Ouro Preto e Mariana (SAMU-OP/MA). Para isso, foram utilizados os dados de atendimento do ano de 2017. Os resultados da implementação do Algoritmo VNS/MHF impactaram positivamente o desempenho do sistema estudado, além de possibilitar a avaliação de cenários alternativos de configuração do sistema (aumento de demanda e diminuição de servidor) e análise de regulamentações específicas (brasileira e internacional). Adicionalmente, o Algoritmo VNS/MHF foi testado utilizando um cenário fictício de maior porte, mostrando a aplicabilidade da ferramenta para tratar problemas de maior escala.

Palavras-chave: Modelo Hipercubo de Filas, *Variable Neighborhood Search*, Localização de Servidores Emergenciais, SAMU.

Abstract

Daily, individuals are subjected to situations in which basic health services need to be requested. An important example of this is SAMU (Emergency Medical Services). For SAMU to maintain or improve its service level, the time between receiving the occurrence and the ambulance's arrival at its destination must be as short as possible. In this work, we developed a tool by embedding the MHF (Hypercube Queuing Model) into the VNS (Variable Neighborhood Search) meta-heuristic, allowing to analyze the particularities of the dispatch policies and server location. The MHF deals with the planning of server-to-customer systems and the choice of VNS is justified by its extensive use in solving various types of location problems. The developed VNS/MHF Algorithm was used to optimize the location of the ambulances of the SAMU in the cities of Ouro Preto and Mariana. For this, service data from 2017 were used. The implementation of the VNS/MHF Algorithm positively impacted the performance of the studied system and enabled the evaluation of alternative system configuration scenarios (increase in demand and decrease in the number of servers) besides the analysis of specific legislation (Brazilian and international). Additionally, the algorithm was tested using a fictitious large-scale scenario, showing the tool's applicability to handle larger-scale problems.

Keywords: Hypercube Queuing Model, Variable Neighborhood Search, Emergency Facilities Location, SAMU.

Lista de Figuras

Figura 1: Representação do Modelo Hipercubo para três servidores 31
Figura 2: Representação do vértice {110} e seus adjacentes 31
Figura 3: Distribuição geográfica da região estudada 42
Figura 4: Processo de atendimento do SAMU 44
Figura 5: Transição da solução inicial para a solução gerada pela função *Shake* 49
Figura 6: Procedimento de Busca Local 49
Figura 7: Mapa da localização e chamados no ano de 2017 54
Figura 8: Divisão da região estudada em átomos geográficos 55
Figura 9: Representação dos estados do SAMU-OP/MA através do Modelo Hipercubo 61
Figura 10: Taxas de ocupação dos servidores nas configurações atual e anterior 67
Figura 11: Taxas de ocupação dos servidores utilizando as políticas de despacho atual e baseada na menor distância 69
Figura 12: Algoritmo VNS/MHF 72
Figura 13: Localização dos servidores na solução final 75
Figura 14: Taxas de ocupação dos servidores utilizando o Algoritmo VNS/MHF 76
Figura 15: Tempo médio de viagem para os cenários de aumento na demanda 77
Figura 16: Taxa de ocupação dos servidores nos cenários de aumento na demanda 78
Figura 17: Localização dos três servidores na solução final 79
Figura 18: Taxas de ocupação dos três servidores 79
Figura 19: Tempos médio de viagem dos servidores no cenário de maior porte 84

Lista de Tabelas

Tabela 1 – Conjuntos utilizados no MALP 22
Tabela 2 – Parâmetros utilizados no MALP 22
Tabela 3 – Variáveis utilizadas no MALP 22
Tabela 4 – Diferenças entre o MALP I e MALP II 24
Tabela 5 – Política de despacho 45
Tabela 6 – Dados analisados 47
Tabela 7 – Parâmetros alterados entre os grupos de instâncias 51
Tabela 8 – Resultados computacionais 53
Tabela 9 – Dados do processo de chegada dos chamados 56
Tabela 10 – Tempo médio de viagem entre os átomos (min) 57
Tabela 11 – Política de despacho dos servidores 58
Tabela 12 – Dados dos tempos de atendimento 59
Tabela 13 – Relação entre tempo médio de serviço e tempo médio de viagem 60
Tabela 14 – Notação utilizada no modelo 61
Tabela 15 – Política de despacho dos servidores na configuração anterior 66
Tabela 16 – Tempos médio de viagem nas configurações atual e anterior do SAMU 67
Tabela 17 – Tempos médio de viagem utilizando as políticas de despacho atual e baseada na menor distância 68
Tabela 18 – Estudos que combinam meta-heurísticas com o MHF 70
Tabela 19 – Tempos médio de viagem utilizando o Algoritmo VNS/MHF 76
Tabela 20 – Soluções encontradas para os cenários de aumento na demanda 77
Tabela 21 – Tempos médio de viagem na análise com três servidores disponíveis 80
Tabela 22 – Tempos médio de viagem para as configurações possíveis de localização do servidor 81
Tabela 23 – Tempos médio de viagem na análise com sete servidores disponíveis 82
Tabela 24 – Tempos médio de viagem no cenário fictício de maior porte 83

Lista de Siglas

AE	Algoritmo Evolucionário
AG	Algoritmo Genético
AHQM	Modelo Hipercubo de Filas Agregado
AMEXCLP	Problema de localização de cobertura máxima esperada ajustado
AQM	Modelo aproximado de filas
BT	Busca Tabu
BVNS	VNS Básico
DSM	Double Standard Model
HC	*Hybridized Hill-climbing*
HQM	Modelo de Localização Hipercubo
LSCP	*Location Set Covering Problem*
MALP	Problema de localização de máxima disponibilidade
MCLP	*Maximal Covering Location Problem*
MEXCLP	Problema de localização de cobertura máxima esperada
MILP	Modelo de programação linear inteira mista
NMEXCLP	Problema de localização de cobertura máxima esperada não-linear
PLSCP	Problema de Cobertura de Conjuntos de Locais Probabilístico
RRAs	Ambulâncias de resposta rápida
RTAs	Ambulâncias de transporte regular
RVNS	*Reduced* VNS
SA	*Simulated Annealing*
SAEs	Serviços de Atendimentos Emergenciais
SAMU	Serviço de Atendimento Móvel de Urgência
SAMU-OP/MA	Serviço de Atendimento Móvel de Urgência de Ouro Preto e Mariana
SQM	*Spatial Queueing Model*
TIMEXCLP	Problema de localização de cobertura máxima esperada com variação de tempo
UPA-MA	Unidade de Pronto Atendimento de Mariana
UPA-OP	Unidade de Pronto Atendimento de Ouro Preto

USA	Unidade de Suporte Avançado
USB	Unidade de Suporte Básico
VNS	*Variable Neighborhood Search*

Sumário

1 Introdução 13

1.1 Contexto Geral 13

1.2 Objetivos do Trabalho 15

1.3 Justificativas/Relevância 16

1.4 Estrutura do Trabalho 17

2 Revisão de Literatura 18

2.1 Sistema de Atendimento Móvel de Urgência – SAMU 18

2.2 Problema de localização de servidores emergenciais 19

2.3 Modelo Hipercubo de Filas 28

2.3.1 Medidas de desempenho do sistema 32

2.4 Otimização com Modelo Hipercubo de Filas 34

2.5 *Variable Neighborhood Search* (VNS) 38

3 Sistema Estudado 42

3.1 Região Estudada 42

3.2 Coleta de Dados 46

4 Validação do *Variable Neighborhood Search* 48

4.1 Meta-heurística Proposta 48

4.2 Modelo Matemático 50

4.3 Instâncias Utilizadas 50

4.4 Comparação entre Modelo Matemático e Meta-heurística 51

5 Aplicação do Modelo Hipercubo de Filas ao SAMU de Ouro Preto e Mariana 54

5.1 Validação das hipóteses do Modelo Hipercubo de Filas 54

5.2 Equações de equilíbrio do sistema 60

5.3 Medidas de desempenho 64

5.4 Resultados da aplicação do Modelo Hipercubo de Filas ao SAMU de Ouro Preto e Mariana 66

5.4.1 Comparação das configurações atual e anterior do sistema 66
5.4.2 Comparação da política atual de despachos com a política de despachos baseada na menor distância 67
6 Abordagem combinando o VNS com o Modelo Hipercubo de Filas 70
6.1 Incorporação do MHF ao VNS 70
6.2 Implementação e resultados do Algoritmo VNS/MHF 74
6.2.1 Análise do SAMU-OP/MA utilizando o Algoritmo VNS/MHF 75
6.2.2 Análise do SAMU-OP/MA considerando aumento na demanda 76
6.2.3 Análise do SAMU-OP/MA utilizando três servidores 78
6.2.4 Análise do SAMU-OP/MA considerando a regulação brasileira 80
6.2.5 Análise do SAMU-OP/MA considerando regulamentações internacionais 81
6.2.6 Análise de cenário fictício de maior porte 83
7 Conclusão e Perspectivas Futuras 85
REFERÊNCIAS 88

1 Introdução

1.1 Contexto Geral

Os serviços de urgência e emergência têm crescido expressivamente no cenário brasileiro e mundial. Tem se observado que vidas podem ser salvas se forem rapidamente socorridas por profissionais bem treinados e qualificados, ainda no ambiente fora dos hospitais e transportadas a um local onde possam receber atendimento com suporte específico para cada caso (MARTINS; PRADO, 2003). Diversos estudos realizados também apontam a importância de avaliação do tempo de resposta e sua relação com a taxa de sobrevida do paciente (BLANCHARD *et al*., 2012; MINAYO; DESLANDES, 2008; MOON *et al*., 2020).

O SAMU é um serviço de saúde que contribui para salvar vidas, garantindo melhores condições para a população. O objetivo do SAMU é chegar rapidamente à vítima após ter ocorrido alguma situação de urgência ou emergência de natureza clínica, cirúrgica, traumática, obstétrica, pediátrica, psiquiátrica, entre outras, que possa levar a sofrimento, sequelas ou mesmo a morte (MINISTÉRIO DA SAÚDE, 2020).

Em serviços médicos emergenciais, a rapidez na resposta ao chamado é um fator crucial para a sobrevida do paciente. O tempo entre a solicitação do serviço e o início do atendimento, denominado tempo de resposta, é uma das medidas mais importantes na análise da qualidade do serviço. Este tempo é dependente de uma série de fatores, tais como a situação do tráfego da região, do local de atendimento, do período do dia, do número de servidores disponíveis naquele momento da chamada (TAKEDA; WIDMER; MORABITO, 2001).

O SAMU-OP/MA é o serviço de atendimento médico emergencial responsável por prestar atendimento nos municípios de Ouro Preto e Mariana. Esse sistema funciona atualmente de forma integrada utilizando 4 servidores (uma USA e três USBs). Uma característica marcante desse sistema são os tempos de deslocamentos dos servidores em função do atendimento em zonas rurais e atendimentos realizados em rodovias. Adicionalmente, foi identificado que a USA, em alguns casos, é utilizada para fazer a transferência de pacientes de Ouro Preto e Mariana para a cidade de Belo Horizonte.

Devido às restrições orçamentárias, o SAMU de Ouro Preto e Mariana (SAMU-OP/MA) não dispõe de grande efetivo, incluindo ambulâncias e tripulações, e, portanto, seus gestores precisam tomar decisões de planejamento e operação, considerando diferentes conflitos entre investimentos, custos operacionais e níveis de serviço oferecidos aos usuários. Uma vez que seu efetivo é limitado, o SAMU-OP/MA diversas vezes acaba levando um tempo considerável, mais de 30 minutos para alguns distritos, para chegar ao local de atendimento. Como já ressaltado anteriormente, o tempo de resposta à solicitação é um fator crucial para a sobrevida da vítima e a minimização de sequelas.

O *Variable Neighborhood Search* (VNS) foi proposto por Mladenović e Hansen (1997) para resolver um problema de roteamento de veículos com *backhauls* (VRPB). Desde então, aplicações dessa meta-heurística estão aumentando rapidamente em número e em áreas de estudo, tais como: teoria da localização (BRIMBERG *et al*., 2016), análise de cluster (COSTA; ALOISE; MLADENOVIĆ, 2017), *scheduling* (ANOKIC *et al*., 2020; FATTAHI *et al*., 2019), roteamento de veículos (ARMAS; MELIÁN-BATISTA, 2015), projeto de rede (MARINAKIS; MIGDALAS; SIFALERAS, 2017), dimensionamento de lotes (MENÉNDEZ *et al*., 2016) e confiabilidade (TODOSIJEVIC *et al*., 2016).

O VNS é uma meta-heurística utilizada para resolver problemas de otimização combinatória cuja ideia básica é uma sistemática mudança de estruturas de vizinhança, tanto dentro de uma fase de descida para encontrar um ótimo local quanto em uma fase de perturbação para sair do vale correspondente a esse ótimo local. Essa meta-heurística possui uma fase estocástica e uma fase determinística, que é representada por uma heurística de busca local.

O Modelo Hipercubo de Filas (MHF) foi introduzido por Larson (1974) para tratar sistemas policiais, porém, desde então, tem sido aplicado em uma variedade de sistemas emergenciais. A literatura apresenta aplicações do MHF em sistemas policiais (MENDES; SANTOS, 2016), bombeiros (RODRIGUEZ; DE LA FUENTE; AGUAYO, 2020), sistemas médicos emergenciais em áreas urbanas (ANSARI; YOON; ALBERT, 2017; TAKEDA; WIDMER; MORABITO, 2007), sistemas médicos emergenciais em rodovias (IANNONI; MORABITO, 2007, 2008; IANNONI; MORABITO; SAYDAM, 2011), reparação de sistemas de ônibus (BOYACI; GEROLIMINIS, 2015), serviço de patrulha em rodovias (GEROLIMINIS; KARLAFTIS; SKABARDONIS, 2009),

sistemas emergências na agroindústria da cana (RODRIGUES et al., 2018; RODRIGUES; MORABITO; SAYDAM, 2016) e sistemas médicos emergenciais em campos de refugiados (BLANK, 2021).

Existem também na literatura trabalhos que incorporam o MHF em modelos de otimização, utilizando diversos métodos de otimização, tais como, Algoritmo Genético (GEROLIMINIS; KEPAPTSOGLOU; KARLAFTIS, 2011; IANNONI; MORABITO; SAYDAM, 2008; SAYDAM; AYTUĞ, 2003), Busca Tabu (TORO-DÍAZ et al., 2015), VNS (BOYACI; GEROLIMINIS, 2015), *Simulated Annealing* (GOLABIAN *et al*., 2021), Algoritmo Iterativo de Otimização (RODRIGUEZ; DE LA FUENTE; AGUAYO, 2020) e Heurística de Substituição de Vértices (BATTA; DOLAN; KRISHNAMURTHY, 1989).

O objetivo principal deste trabalho é estudar instalações emergenciais de saúde, focando no problema de localização de servidores emergenciais, utilizando métodos de otimização e teoria de filas. Nesse caso, pretende-se aproveitar as vantagens de cada ferramenta utilizada, a análise descritiva dos modelos de filas e a otimização para testar as configurações geradas.

Este trabalho está localizado dentro do enorme campo de estudo das instalações de saúde. As instalações emergenciais são um subconjunto das de saúde, sendo a localização de servidores parte essencial destas (AHMADI-JAVID; SEYEDI; SYAM, 2017).

O presente estudo trata-se de uma pesquisa quantitativa, sendo classificada como pesquisa axiomática normativa. Desenvolve normas, políticas, estratégias e ações, a fim de melhorar os resultados disponíveis na literatura, encontrar uma solução ótima para um problema novo ou comparar o desempenho de estratégias que tratam um mesmo problema (MIGUEL *et al*., 2018). De acordo com Miguel *et al*. (2018), essa classe baseia-se em modelos que prescrevem uma decisão para o problema – em geral, modelos de programação matemática.

1.2 Objetivos do Trabalho

O objetivo geral deste trabalho é desenvolver uma abordagem que incorpore o modelo Hipercubo de filas à meta-heurística VNS, permitindo analisar as particularidades

das políticas de despacho e localização dos servidores. Pretende-se também mostrar que essa abordagem pode ser utilizada para analisar o SAMU de Ouro Preto e Mariana.

Esse objetivo geral pode ser desmembrado nos seguintes objetivos específicos:

- Estender a meta-heurística VNS para tratar o problema proposto;
- Adaptar e aplicar o modelo Hipercubo de filas ao SAMU de Ouro Preto e Mariana;
- Incorporar o modelo Hipercubo de filas à meta-heurística VNS;
- Utilizar a abordagem desenvolvida para calcular e analisar os indicadores de desempenho para o sistema;
- Utilizar a abordagem desenvolvida para explorar cenários, avaliando configurações alternativas do sistema (aumento de demanda e diminuição de servidor) e analisando regulamentações específicas (brasileira e internacional);
- Testar a abordagem desenvolvida utilizando cenário fictício de maior escala de modo a permitir a avalição do potencial de análise dessa abordagem em situações realista enfrentadas por sistemas emergenciais.

1.3 Justificativas/Relevância

É esperado que o SAMU responda com eficiência e eficácia as demandas de urgências e emergências de toda população, dispondo de competência técnico-científica, ética, educação pública, recursos, tecnologias e processos de avaliação constantes. O SAMU é um ponto chave do sistema de saúde moderno, sendo responsável pelo componente pré-hospitalar, consistindo em atendimento médico e atividades de transporte, desde a chegada de uma chamada de emergência até a liberação do paciente ou sua transferência para um hospital. Desta forma, o SAMU possui uma importante função no sistema de saúde moderno e, sua habilidade de responder de forma eficiente às chamadas pode ter um impacto significativo na saúde do paciente e em sua recuperação (BÉLANGER; RUIZ; SORIANO, 2019).

Tendo em vista o panorama da necessidade de avaliação e estudos constantes voltados para a análise do desempenho e melhorias dos sistemas emergenciais e da integração proposta anteriormente pelos autores citados, o presente estudo destaca-se com grande relevância ao propor um estudo minucioso do funcionamento do SAMU-OP/MA em vigência, além de sugerir melhorias e avanços para o sistema.

Do ponto de vista científico, esse estudo incorpora o MHF ao VNS, visando otimizar a localização das ambulâncias no SAMU-OP/MA, considerando as particularidades da política de despacho de servidores.

Foi identificado na literatura, apenas um estudo que utiliza o MHF incorporado ao VNS (BOYACI; GEROLIMINIS, 2015). Ainda assim, nesse estudo, os autores buscam minimizar os serviços interdistritais. Dessa forma, esse trabalho traz uma importante contribuição para a literatura.

Adicionalmente, a ferramenta desenvolvida permite melhorar o desempenho do sistema com a melhor utilização de recursos escassos, uma vez que possibilita a otimização da configuração do sistema, sem necessariamente sugerir investimentos adicionais.

1.4 Estrutura do Trabalho

Este trabalho está dividido em seis capítulos, sendo o primeiro uma apresentação da revisão bibliográfica dando fundamento teórico à pesquisa. O segundo apresenta o sistema estudado. No terceiro, é apresentada a validação do VNS. O quarto descreve a aplicação do MHF ao SAMU-OP/MA. No quinto, é desenvolvida a abordagem combinando o VNS com o MHF. E, o último apresenta as conclusões e pesquisas futuras.

2 Revisão de Literatura

Neste capítulo é realizada uma revisão de literatura dos temas estudados neste trabalho. São eles, SAMU, problemas de localização de servidores emergências, Modelo Hipercubo de Filas, métodos que mesclam otimização com o Modelo Hipercubo e VNS.

2.1 Sistema de Atendimento Móvel de Urgência – SAMU

A política nacional de atenção às urgências (MINISTÉRIO DA SAÚDE, 2003a), fruto do trabalho em parceria do Ministério da Saúde, Conselho Nacional de Saúde, estados e municípios visa promover a saúde de forma solidária, ágil e eficiente. Dentro deste contexto, institui-se o componente pré-hospitalar móvel, por intermédio da implantação do SAMU, através da Portaria nº 1.864/GM, de 29 de setembro de 2003, em todo território brasileiro, no âmbito do Sistema Único de Saúde (SUS) (MINISTÉRIO DA SAÚDE, 2003b).

O objetivo principal do SAMU é realizar o atendimento à vítima no menor tempo possível. As unidades móveis podem ser ambulâncias, motolâncias, aeromédicos ou ambulanchas de acordo com a disponibilidade e necessidade de cada situação. De modo a otimizar o tempo de resposta dos chamados da população e o encaminhamento aos serviços hospitalares, as ambulâncias do SAMU são distribuídas estrategicamente, com o intuito de prestar o atendimento à vítima no menor tempo possível.

O SAMU conta com ambulâncias de suporte avançado (USA) e de suporte básico (USB) que visam o estabelecimento da estruturação de uma rede de serviços hierarquizada e regionalizada para a atenção às urgências, além da implantação de um processo de regulação que promova eficiência cotidiana ao sistema (MINISTÉRIO DA SAÚDE, 2003a). Segundo a Portaria nº 1010 de 21 de maio de 2012 (MINISTÉRIO DA SAÚDE, 2012), a USB deve ser tripulada por no mínimo dois profissionais, sendo um condutor de veículo de urgência e um técnico ou auxiliar de enfermagem. Já a USA é tripulada por no mínimo três profissionais, sendo um condutor de veículo de urgência, um enfermeiro e um médico.

O atendimento do SAMU inicia-se a partir do contato telefônico gratuito, no número 192, da vítima ou de alguém próximo que esteja a assistindo. Neste primeiro momento são prestadas orientações básicas fornecidas por um técnico que identifica a

emergência e coleta as primeiras informações sobre a vítima e sua localização. A seguir, a chamada é repassada para o médico regulador, que fornecerá instruções mais específicas e acionará as ambulâncias quando necessário (MINISTÉRIO DA SAÚDE, 2020). Quando o médico regulador identifica a necessidade de enviar uma ambulância para realizar o atendimento, baseado no relato da vítima ou de alguém próximo a ela, ele decide entre o envio de uma USA ou uma USB. Ao chegar no local do chamado, a tripulação da ambulância presta os primeiros socorros à vítima e dado a complexidade do caso clínico, realiza a remoção dela para um centro médico. Após deixar a vítima sob os cuidados do centro médico, caso não tenha ocorrido outra chamada no sistema, a ambulância retorna para base, para repor alguns materiais utilizados durante o atendimento e, caso necessário, passa por um processo de limpeza, onde fica aguardando o próximo chamado.

A fim de ampliar o atendimento do SAMU às populações dos municípios em todo o território nacional, o Ministério da Saúde, por meio da Portaria nº 1010 de 21 de maio de 2012 (MINISTÉRIO DA SAÚDE, 2012), Capítulo III, Art. 7º estabelece que o componente SAMU deverá ser regionalizado, atendendo as diretrizes e parâmetros técnicos definidos pela presente portaria. Entretanto, esse processo de regionalização encontra-se em diferentes estágios, no qual observa-se desde processos concluídos até não inicializados.

2.2 Problema de localização de servidores emergenciais

O objetivo dos serviços de atendimentos emergenciais (SAE) é reduzir mortalidade, incapacidade e sofrimento nas pessoas. Um meio de se atingir esse objetivo é diminuir o tempo de resposta dos chamados. Para alcançar esse objetivo, é necessário que os servidores emergenciais sejam localizados estrategicamente.

A literatura em problemas de localização de servidores emergenciais (PLSE) é rica e diversa. Ao longo dessa revisão bibliográfica é feito um relato do desenvolvimento de modelos para o PLSE. Desde o surgimento dos modelos mais simples, na década de 1970, que eram determinísticos e simplificados até os estocásticos atuais, que são mais robustos e representam melhor a realidade.

De acordo com Li *et al.* (2011), os modelos para o PLSE podem ser divididos em três categorias: de cobertura, que visam otimizar a cobertura para chamadas de emergência com uma distância predefinida; de *p*-medianas, que minimizam a distância

média de serviços entre todos os pontos de demanda; e de *p*-centro, que minimizam a distância máxima de serviço entre todos os pontos de demanda.

Além das três categorias de modelos de localização citadas acima, estes também podem ser divididos em determinísticos e estocásticos. Os modelos determinísticos não consideram que os parâmetros do sistema sofrem variabilidade, por exemplo, considera o tempo de deslocamento e a probabilidade de o servidor estar ocupado como valores fixos. Já os modelos estocásticos assumem que alguns ou todos os parâmetros do sistema sofrem variações, quer seja ao longo do dia, meses ou anos.

No início dos anos 1970, dois modelos determinísticos de cobertura foram estruturados e resolvidos:

1. *Location Set Covering Problem* (LSCP) – desenvolvido por Toregas *et al*. (1971), esse modelo busca posicionar o menor número de servidores garantindo que todos os pontos de demanda tenham pelo menos um servidor que consiga atendê-lo com um tempo de resposta ou distância padronizado.
2. *Maximal Covering Location Problem* (MCLP) – esse modelo aloca um número limitado de facilidades de forma que a população assistida por pelo menos um servidor, que atenda o tempo de viagem ou a distância padronizada, seja maximizada (CHURCH; REVELLE, 1974).

Dadas as limitações orçamentárias que os serviços emergenciais enfrentam, o número limitado de servidores do modelo MCLP, geralmente, é insuficiente para cobrir todas as áreas de demanda dentro do tempo ou distância padrão. Quando esses dois modelos determinísticos de cobertura são aplicados a sistemas de emergência que apresentam uma alta taxa de chegada de chamadas em relação ao número de servidores, os mesmos podem levar a soluções que não atendam ao conjunto original de expectativas de performance.

A preocupação de fornecer um serviço de confiança dentro de um tempo (ou distância) padrão, conduziu à uma variedade de modelos que levam em consideração o efeito do congestionamento. Na esperança de que a cobertura múltipla de áreas de demanda aumentaria a confiabilidade, Daskin e Stern (1981) e Berlin e Liebman (1974) buscaram maximizar a cobertura redundante, dado um número fixo de facilidades. Usando uma simples modificação no LSCP, o conjunto de possíveis localidades de

alocação das ambulâncias não é necessariamente igual ao conjunto de todas as regiões de demanda. Champan e White (1974) formularam uma versão probabilística do LSCP sob a suposição de um sistema com taxa de ocupação uniforme para todos os servidores.

O problema de localização de cobertura máxima esperada (MEXCLP) foi desenvolvido por Daskin (1982) utilizando a mesma suposição de Champan e White (1974). O MEXCLP visa maximizar a cobertura, levando em conta que os servidores podem estar ocupados quando chegam as chamadas no sistema. Basicamente, o MEXCLP pode ser estabelecido como: alocar *M* facilidades em possíveis lugares em uma rede que maximiza a população esperada a ser coberta, dadas uma distância (tempo) padrão *S* e uma probabilidade de cada veículo estar ocupado ser igual a *p*. Os resultados de Daskin (1982) demonstram que modelos determinísticos subestimam significativamente o número de veículos necessários para cobrir uma fração de demanda pré-especificada.

O problema de localização de máxima disponibilidade (MALP), desenvolvido por Revelle e Hogan (1989), visa localizar um determinado número de servidores tal que a máxima população possível de uma dada região seja coberta a menos de uma distância crítica *S* pré-definida. Sendo assim, um ponto é tido como "coberto" se pelo menos um servidor estiver localizado em uma região onde consiga chegar a esse ponto em um tempo menor ou igual ao tempo padrão, dado um fator de confiança α. O MALP foi formulado em duas versões. A primeira (MALP I) utiliza a suposição de que a taxa de ocupação dos servidores no sistema é igual. Já a segunda (MALP II) assume diferentes taxas de ocupação para os servidores, geralmente utilizando o Modelo Hipercubo de Filas de Larson (1974) para determinar as diferentes taxas de ocupação dos servidores (REVELLE; HOGAN, 1989).

Os conjuntos, parâmetros e variáveis de decisão do MALP I estão representados nas Tabelas 1, 2 e 3, respectivamente. O cálculo do parâmetro *b*, apresentado na Tabela 2, determina o número de servidores necessários para que a confiabilidade seja maior ou igual a α, isto é, cada área de demanda $i \in I$ requer pelo menos *b* servidores disponíveis dentro do raio de cobertura *S* para ser considerada coberta, com confiabilidade α (GALVÃO; MORABITO, 2008).

Tabela 1 – Conjuntos utilizados no MALP

Conjunto	Descrição
I	Regiões de demanda
J	Regiões candidatas
$N_i = \{j \mid t_{ij} \leq S\}$	Regiões candidatas dentro do raio de cobertura da região de demanda i

Tabela 2 – Parâmetros utilizados no MALP

Parâmetro	Descrição
t_{medio}	Tempo médio de atendimento
S	Raio de cobertura
f_i	Número de chamadas por dia da região i
t_{ij}	Menor distância da região j para a área de demanda i
p	Número de servidores
α	Confiabilidade
$q = t_{medio} * \sum_{i \in I} \frac{f_i}{24P}$	Taxa de ocupação dos servidores
$b = \left\lceil \frac{\log(1-\alpha)}{\log q} \right\rceil$	Servidores disponíveis a menos da distância crítica S com confiabilidade α

Tabela 3 – Variáveis utilizadas no MALP

Variável	Descrição
x_j	Variável binária que assume valor 0 se um servidor não é posicionado na região j e valor 1 se um servidor é posicionado nessa região
y_{ik}	Variável binária que assume valor 1 se k servidores estão alocados dentro do raio de cobertura da região de demanda i, e o valor 0, caso contrário

A formulação matemática do MALP I é apresentada a seguir:

$$\max z = \sum_{i \in I} f_i y_{ib} \qquad (1)$$

Sujeito a

$$\sum_{k=1}^{b} y_{ik} \leq \sum_{j \in N_i} x_j \qquad \forall i \in I \qquad (2)$$

$$y_{ik} \leq y_{ik-1} \qquad \begin{cases} \forall i \in I \\ k = 2, \dots, b \end{cases} \qquad (3)$$

$$\sum_{j \in J} x_j = p \qquad (4)$$

$$x_j, y_{ik} \in \{0,1\} \qquad i \in I, j \in J, k = 1, \dots, b \qquad (5)$$

A função objetivo (1) maximiza o número de chamadas atendidas por dia, com nível de confiabilidade α. As restrições (2) garantem que cada região de demanda $i \in I$ é

coberta com confiabilidade α se existirem pelo menos b servidores dentro de S. Já as restrições (3) mostram que para uma região de demanda ser coberta por k servidores, ela tem que ser coberta por pelo menos $k-1$ servidores. A restrição (4) garante que p servidores devem ser alocados. As restrições (5) definem a natureza binária das variáveis de decisão.

No MALP II é relaxada a suposição de que a taxa de ocupação de todos os servidores é igual e permite que a taxa de ocupação seja diferente dentro da região analisada. Para isso, incorpora-se o modelo Hipercubo (LARSON, 1974) que permite calcular as diferentes taxas de ocupação para os servidores alocados em diferentes localidades da região estudada. A formulação do MALP II define r_j como a taxa de ocupação de um veículo na localidade j. A restrição da probabilidade de um veículo estar disponível para atender a região de demanda i dentro do tempo padrão S com uma confiabilidade de pelo menos α é dada por:

$$1 - \prod_{j \in N_i} r_j^{x_j} = \alpha \tag{6}$$

Que pode ser linearizado, tomando logaritmos:

$$\sum_{j \in N_i} (\log r_j) x_j \leq \log(1 - \alpha) \tag{7}$$

Se as restrições de probabilidade forem utilizadas, a solução irá requerer o uso de um código zero-um eficiente, capaz de resolver problemas de grande porte. Sendo assim, Revelle e Hogan (1989) desenvolveram uma metodologia, que utiliza taxas específicas de ocupação para as áreas.

Na metodologia de Revelle e Hogan (1989), a taxa média de ocupação dos servidores ao redor da região de demanda i pode ser estimada por:

$$q_i = \frac{\bar{t} * \sum_{k \in M_i} f_k}{24 \sum_{k \in M_i} x_j} = \frac{F_i}{\sum_{j \in N_i} x_j} \tag{8}$$

Onde:

$$F_i = \frac{\left(t * \sum_{k \in M_i} f_k\right)}{24} \tag{9}$$

e M_i é o conjunto de pontos de demanda dentro do raio de cobertura. O número de veículos em N_i deve ser maior ou igual ao menor inteiro que satisfaça a restrição não-linear:

$$\sum_{j \in N_i} x_j \geq b_i \tag{10}$$

Onde b_i é o menor inteiro que satisfaz:

$$1 - (f_i b_i)^{b_i} \geq \alpha \tag{11}$$

Então, a restrição não-linear pode ser transformada em uma forma linear, satisfazendo a restrição original.

A diferença entre o MALP I e o MALP II está na forma como a taxa de ocupação é estimada. No MALP I, a taxa de ocupação é assumida como uniforme no sistema e fornecer *b* servidores para cada área de demanda satisfaz as restrições de chance com a confiabilidade desejada. Assim, maximizar a população coberta com confiabilidade α implica em maximizar a população coberta por *b* ou mais servidores. No MALP II, as taxas de ocupação são estimadas nas regiões definidas por cada uma das restrições de chance. Dessa forma, no MALP II, maximizar a população coberta com confiabilidade α implica em maximizar a demanda coberta por seu específico b_i, conforme sumarizado na Tabela 4.

Tabela 4 – Diferenças entre o MALP I e MALP II

MALP I	**MALP II**
Considera que a taxa de ocupação dos servidores é uniforme, ou seja, igual para todos os servidores do sistema.	Permite que a taxa de ocupação seja diferente nas várias seções da cidade.
A taxa média de ocupação dos servidores é dada por: $q = t_{medio} * \sum_{i \in I} \frac{f_i}{24p}$	A taxa média de ocupação dos servidores ao redor da região de demanda *i* é dada por: $q_i = \frac{\bar{t} * \sum_{k \in M_i} f_k}{24 \sum_{k \in M_i} x_j} = \frac{F_i}{\sum_{j \in N_i} x_j}$
Servidores disponíveis a menos da distância crítica *S* com confiabilidade α: $b = \left\lceil \frac{\log(1-\alpha)}{\log q} \right\rceil$	Servidores disponíveis a menos da distância crítica *S* com confiabilidade α: $1 - (f_i b_i)^{b_i} \geq \alpha$
Restrição linear que considera a região de demanda $i \in I$ coberta se: $\sum_{k=1}^{b} y_{ik} \leq \sum_{j \in N_i} x_j$	Restrição que considera a região de demanda $i \in I$ coberta se: $\sum_{k=1}^{b_i} y_{ik} \leq \sum_{j \in N_i} x_j$

Saydam, Repede e Burwell (1994) realizaram um estudo comparativo dos modelos de localização de servidores já existentes na literatura. Segundo os autores, uma vez que a cobertura é relacionada ao número de veículos da frota, o planejador do SAE deve tentar atingir um balanço entre cobertura e orçamento. Os autores descrevem que existem dois tipos de PLSE: o primeiro busca um conjunto de localizações que maximiza a cobertura, dada uma frota existente. O segundo procura encontrar o menor número de veículos (e suas localizações) para atingir um nível mínimo de cobertura definido. Saydam, Repede e Burwell (1994) concluíram que para taxas de ocupação baixas, o AMEXCLP, o NMEXCLP e o TIMEXCLP tendem a superestimar o valor da cobertura. Já para taxas de ocupação elevadas, esses modelos subestimam esse valor.

Brotcorne, Laporte e Semet (2003) traçaram a evolução dos modelos de localização e realocação de servidores emergenciais ao longo dos últimos 30 anos e classificaram os modelos em duas categorias principais: determinísticos e probabilísticos. Além disso, citam que modelos dinâmicos têm sido desenvolvidos para realocar as ambulâncias ao longo do dia. Os autores sugerem que qualquer desenvolvimento na área de localização de servidores emergenciais provavelmente irá explorar a área de modelos de realocação dinâmica junto com rápidas e precisas heurísticas.

Rajagopalan *et al*. (2007) aplicam quatro meta-heurísticas para o MEXCLP, Algoritmo Evolucionário (AE), Busca Tabu (BT), *Simulated Annealing* (SA) e *Hybridized Hill-climbing* (HC). Os resultados obtidos pelas quatro meta-heurísticas e pela resolução do modelo matemático através do CPLEX foram comparados entre si, utilizando um estudo comparativo através do planejamento de experimentos. Com exceção do AE, as meta-heurísticas superaram significativamente o CPLEX, conforme o tamanho do problema vai aumentando, tanto em tempos médio, quanto nos desvios padrão. Além disso, a BT e o SA encontraram as melhores soluções em menor tempo, em comparação com as outras meta-heurísticas e com as resoluções utilizando o CPLEX. Dessa forma, essas duas meta-heurísticas foram identificadas como bons métodos de resolução de modelos similares ao MEXCLP, para problemas de larga escala.

Rajagopalan e Saydam (2009) propuseram uma nova abordagem que determina as localizações, dado um número fixo de ambulâncias, com o objetivo de minimizar as distâncias/tempos de resposta, enquanto atende os requisitos de cobertura. Os modelos formulados pelos autores são baseados no modelo p-mediana clássico de Hakimi (1965)

e inspirado no MALP, e no modelo p-centro de Revelle e Hogan (1989). Os autores utilizam um procedimento BT reativo e uma heurística de busca gulosa para resolver os modelos propostos.

Mayorga, Bandara e Mclay (2013) apresentaram um estudo de integração das políticas de despacho com divisão dos distritos para melhorar o desempenho dos sistemas médicos emergenciais. Os autores mensuraram o desempenho em termos da probabilidade de sobrevivência do paciente. São propostas várias políticas de despacho, que são fornecidas como insumos para um modelo de simulação que compara o desempenho de diferentes políticas. As áreas de resposta, ou distritos, são projetadas usando uma heurística construtiva que considera a cobertura esperada ajustada. As políticas de despacho intra e interdistrital são desenvolvidas considerando o grau de urgência da chamada. Os resultados computacionais mostraram que políticas integradas de despacho e divisão em distritos são vitais para aumentar as chances de sobrevivência dos pacientes.

Grande parte do trabalho anterior em despacho médico emergencial concentrou-se em localizar as ambulâncias, na dinâmica das filas ou no envio de ambulância. Em contraste com trabalhos anteriores na área, Ansari, Mclay e Mayorga (2015) investigaram como integrar essas três questões em um único modelo que localiza ambulâncias enquanto determina como usá-las, ao mesmo tempo que mantém o realismo através de tempos de viagem estocásticos. Com o intuito de determinar como alocar e despachar ambulâncias, através do *design* de distritos, os autores propuseram um modelo de programação linear inteira mista (PLIM). O modelo permite incerteza tanto no tempo de viagem quanto na disponibilidade das ambulâncias e visa maximizar o nível de cobertura. As estações onde as ambulâncias devem ser alocadas e o despacho delas, de acordo com uma lista preferencial, são determinadas pelo modelo proposto.

Barneveld, Mei e Bhulai (2017) utilizaram tabelas de conformidade, que indicam as localizações desejadas das ambulâncias disponíveis, para um sistema médico emergencial holandês que apresenta dois tipos de veículos: ambulâncias de resposta rápida (RRAs) e ambulâncias de transporte regular (RTAs). A principal diferença entre os dois tipos é que as RRAs são mais rápidas, mas não podem ser utilizadas para transportar um paciente para o hospital. A ideia desse sistema médico emergencial é de despachar uma RRA para atender a chamada, desde que ela consiga chegar ao paciente

dentro do tempo padrão estipulado. Chegando ao local, o servidor presta os primeiros socorros e identifica se a vítima precisa ser removida para um hospital, se sim, a RTA mais próxima é enviada para realizar a remoção. Os autores expandem o modelo MEXCLP2, proposto por Mclay (2009), para ajustar na estrutura de tabelas de conformidade.

Trujillo *et al*. (2020) conduziram um estudo para determinar os benefícios, custos, vantagens e desvantagens de usar diferentes estratégias de realocação de ambulâncias, quando ocorrem variações na demanda ao longo do tempo, em Tijuana, México. Foram avaliadas três estratégias, todas baseadas no *Double Standard Model* (DSM). A primeira abordagem aplica uma versão robusta do DSM para encontrar as melhores soluções de *trade-off* em todos os períodos ou cenários possíveis ao longo do dia. A segunda abordagem aplica o DSM a cada cenário independente~~s~~ e executa realocações com base nas configurações das ambulâncias em diferentes cenários. A terceira abordagem consiste em utilizar um modelo de realocação explícito, o DSM de vários períodos. As abordagens são avaliadas com base no percentual de cobertura dupla, número total de realocações, tempo de viagem para realocação, distância de viagem para realocação e custo financeiro da realização das realocações.

De acordo com Yoon e Albert (2020), alguns sistemas de serviços médicos de emergência equipados com veículos de emergência avançados e básicos costumam despachar os dois tipos de veículos para uma chamada, denominada resposta múltipla. Para avaliar o valor da resposta múltipla, os autores formularam um modelo utilizando processo de decisão de Markov que determina dinamicamente que tipo de veículo (*s*) enviar para atender as chamadas. Os autores mostraram que as políticas ideais são separáveis por classe e por meio de experimentos numéricos demonstram que a resposta múltipla pode melhorar significativamente o desempenho do sistema quando as necessidades de saúde dos pacientes são incertas.

Sudtachat *et al*. (2020) estenderam o modelo de conformidade aninhada (*Nested-compliance Model* - NCM), que fornece uma política de realocação para considerar um limite superior no tempo de realocação, particionando toda a região em subáreas menores (distritos). A decisão sobre como particionar a área de serviço em subáreas menores, nas quais cada subárea opera independentemente, é analisada sob estratégias de realocação separadas. Depois que a solução distrital é determinada, BT é usada para alocar estações

e zonas aos distritos e a solução ideal de NCM é aplicada a cada subárea. O objetivo foi maximizar a cobertura geral esperada realizada. O desempenho das políticas combinadas de realocação e dimensionamento de áreas foi comparado a uma política estática (sem realocação e sem dimensionamento de áreas) com base no problema de localização de cobertura máxima esperada ajustado (AMEXCLP) em um sistema simulado usando dados do mundo real. Os resultados numéricos mostraram os benefícios do modelo proposto em relação ao AMEXCLP.

2.3 Modelo Hipercubo de Filas

O Modelo Hipercubo de Filas (MHF) é utilizado para avaliar os sistemas onde os servidores se deslocam para fornecer algum tipo de serviço para o cliente (*server-to-customer*). Trata-se de um modelo analítico que expande o espaço de estados de um sistema de fila M/M/m, sistema que possui *m* servidores e as chegadas são determinadas por um processo de Poisson e os tempos de serviço têm uma distribuição exponencial, a fim de representar cada servidor individualmente e, desta maneira, considerar políticas de despacho de servidores no sistema.

O MHF, a princípio, foi elaborado para lidar com sistemas policiais, porém, desde então, tem sido aplicado para uma variedade de sistemas emergenciais. A literatura contém aplicações do MHF em sistemas policiais (LARSON; MCKNEW, 1982; MENDES; SANTOS, 2016), bombeiros (RODRIGUEZ; DE LA FUENTE; AGUAYO, 2020; SWERSEY; GOLDRING; GEYER, 1994), sistemas médicos emergenciais em áreas urbanas (ANSARI; YOON; ALBERT, 2017; BURWELL; MCKNEW; JARVIS, 1992; DE SOUZA *et al*., 2015; RAJAGOPALAN; SAYDAM; XIAO, 2008; TAKEDA; WIDMER; MORABITO, 2007; TORO-DÍAZ *et al*., 2013, 2015), sistemas médicos emergenciais em rodovias (IANNONI; MORABITO, 2007, 2008; IANNONI; MORABITO; SAYDAM, 2011; MENDONÇA; MORABITO, 2001; MOBIN; LI; AMIRI, 2015), reparação de sistemas de ônibus (BOYACI; GEROLIMINIS, 2015; GEROLIMINIS; KEPAPTSOGLOU; KARLAFTIS, 2011), serviço de patrulha em rodovias (GEROLIMINIS; KARLAFTIS; SKABARDONIS, 2009), sistemas emergências na agroindústria da cana (RODRIGUES *et al*., 2018; RODRIGUES; MORABITO; SAYDAM, 2016) e sistemas médicos emergenciais em campos de refugiados (BLANK, 2021).

A aplicação do MHF está estruturada com base em noves hipóteses (LARSON; ODONI, 2007), que são:

I. Átomos geográficos: a região estudada é dividida em N_A átomos geográficos, onde cada um se refere a uma fonte de chamadas independente, levando em consideração as complexidades temporais (trânsito) e geográficas (condições da via) de cada átomo.

II. Processo de chegada independente: as chamadas emergenciais de cada átomo j (j = 1, 2, ..., N_A) são concebidas segundo o processo de *Poisson*. A taxa de chegada λ_j é conhecida ou pode ser estimada.

III. Tempo de viagem: tempo de deslocamento entre um átomo *i* para um átomo *j* (i, j = 1, 2, ..., N_A) e pode ser conhecido ou estimado utilizando-se de conceitos de probabilidade geométrica.

IV. Servidores: existem *N* servidores no sistema, homogêneos ou não, que estão distribuídos no espaço delimitado pelo sistema analisado, podendo migrar para qualquer átomo.

V. Localização dos servidores: se um servidor se encontra livre, ele pode estar localizado de forma estacionária ou se movimentando, sendo sua localização conhecida ao menos probabilisticamente, ou seja, é necessário conhecer a região na qual ele se movimenta.

VI. Despacho do servidor: em uma chamada emergencial, será designado apenas um servidor para aquele atendimento, mas se ocorrer de todos estarem ocupados a chamada irá aguardar em uma fila.

VII. Preferências de despacho: existe uma ordem de preferências de despacho para cada chamada realizada, dependendo de em qual átomo a chamada ocorreu.

VIII. Tempo de atendimento: o tempo de atendimento a uma chamada é conhecido e inclui o tempo de *setup* (preparação), tempo de viagem do servidor ao lugar de origem da chamada, o tempo em cena e o tempo de retorno a base. Os servidores, em sua maioria, possuem tempos de atendimento distintos.

IX. Dependência do tempo de serviço com o tempo de viagem: as variações ocorridas nos tempos de atendimento ocorrem devido a variações no tempo gasto na viagem do servidor ao local da chamada, mas são considerados de segunda ordem (em comparação com as variações do tempo em cena e de preparação).

Muitos estudos propõem extensões para o MHF, relaxando uma ou mais das suas principais suposições. Por exemplo, os trabalhos de Chelst e Barlach (1981), Iannoni (2005), Iannoni e Morabito (2007) e Iannoni, Morabito e Saydam (2008) consideram o despacho de múltiplos servidores. Larson e Mcnkew (1982), Atkinson *et al*. (2008) e Boyaci e Geroliminas (2015) incluem um terceiro estado para os servidores. Morabito, Chiyoshi e Galvão (2008) e Geroliminas, Karlaftis e Skabardonis (2009) assumem que os servidores são não homogêneos. Iannoni, Morabito e Saydam (2009), Akinson *et al*. (2008) e Rodrigues *et al*. (2017) endereçam o *backup* parcial de servidores. Burwell, Jarvis e Mcknew (1993), Takeda, Widmer e Morabito (2007), Iannoni e Morabito (2007), De Souza *et al*. (2015) e Barneveld, Mei e Bhulai (2017) levam em consideração que diferentes servidores atendem diferentes tipos de chamada na mesma região. De Souza et al. (2015) e Rodrigues *et al*. (2017) endereçam explicitamente políticas de prioridade no serviço de atendimento dos clientes em fila.

Um sistema possui dois estados de funcionamento para seus servidores, livre e ocupado. Quando o servidor está livre, ele assume o valor 0, já, quando está ocupado, recebe valor 1, por isso o sistema possui 2^n estados. Por exemplo, para um sistema composto de três servidores, tem-se 8 estados ($\{000\}$, $\{001\}$, $\{010\}$, ..., $\{111\}$). A Figura 1 representa o espaço de estados para um sistema com três servidores.

Na Figura 1 é possível perceber que a mudança de um estado para o outro se dá através das arestas, uma vez que em um determinado instante apenas um servidor modifica seu estado de livre para ocupado ou vice-versa. Considerando um sistema com fila infinita, a cauda formada pelos estados $S_4, S_5, etc.$, representa os estados em que os servidores se encontram ocupados e existem chamados aguardando para serem atendidos, ocasionando assim uma fila no sistema.

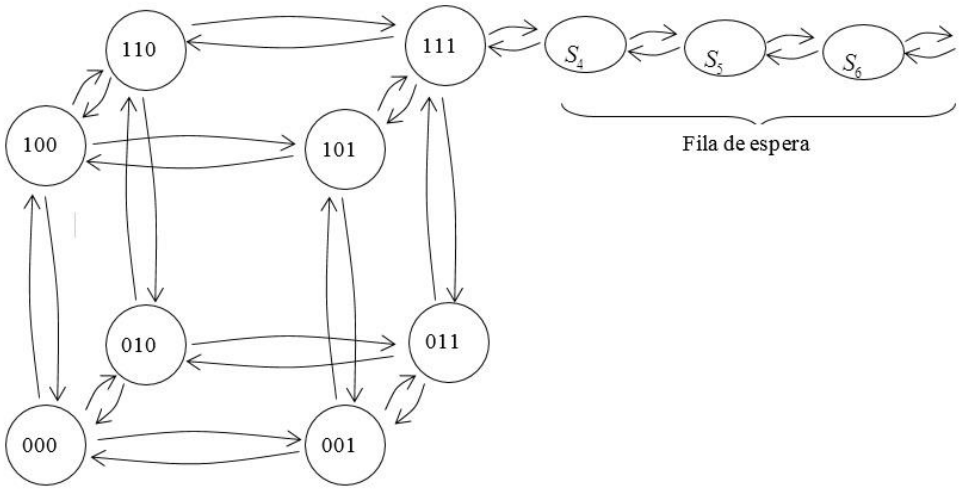

Figura 1: Representação do Modelo Hipercubo para três servidores

Por exemplo, {110} é um estado intermediário em que os servidores 1 e 2 estão atendendo às chamadas e o servidor 3 está disponível. O término do atendimento por parte de um dos dois servidores ocupados e a chegada de mais um chamado ao sistema, são as transições possíveis a partir deste estado. A Figura 2 ilustra as transições possíveis a partir do estado em questão e a equação de equilíbrio (12) indica estas mudanças.

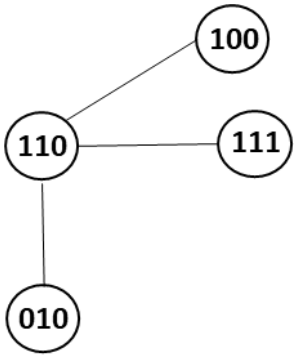

Figura 2: Representação do vértice {110} e seus adjacentes

$$(\lambda + \mu_1 + \mu_2)P_{110} = \lambda P_{100} + \lambda P_{010} + \mu_3 P_{111} \qquad (12)$$

As transições para fora do estado {110} são retratadas pelo lado esquerdo da Equação (12):

i. {110} para {111}: chega uma chamada a partir de qualquer átomo do sistema, taxa de chegada λ, fazendo assim com que todos os servidores do sistema fiquem ocupados;

ii. {110} para {100}: ocorre o término do serviço por parte do servidor 2, com taxa μ_2;

iii. {110} para {010}: acontece o fim da atividade do servidor 1, com taxa μ_1.

O fluxo das transições para dentro do estado {110} são retratadas pelo lado direito da Equação (12):

i. {100} para {110}: ocorre a chegada de uma chamada e o servidor 2 se desloca para realizar seu atendimento;

ii. {010} para {110}: acontece uma chamada que faz com que o servidor 1 seja delegado para o serviço;

iii. {111} para {110}: ocorre o término do serviço do servidor 3, com taxa μ_3.

Uma vez que as equações de equilíbrio foram desenvolvidas, um sistema de equações lineares indeterminado é obtido. Portanto, faz-se necessário acrescentar uma equação adicional para determinar as probabilidades dos estados do MHF. Em sistemas que admitem fila com capacidade infinita, a equação de normalização é dada por (13), em que a taxa de ocupação média do sistema (ρ) é dada por $\rho = \frac{\lambda}{\mu}$.

$$P_{000} + P_{100} + P_{010} + P_{001} + P_{110} \ldots + \frac{P_{111}}{\rho} = 1 \qquad (13)$$

2.3.1 Medidas de desempenho do sistema

A partir das probabilidades de equilíbrio dos estados do MHF, torna-se possível a utilização destas para se obter medidas de desempenho para o sistema. Muitas medidas podem ser estimadas, seja em toda região, para cada servidor ou por região.

A taxa de ocupação, ou carga de trabalho, de um servidor *i* (ρ_i), é a fração do tempo que este servidor está ocupado, que é a soma das probabilidades de equilíbrio dos estados nos quais este servidor está ocupado. Portanto, a carga de trabalho do servidor *n* é representada pela Equação (14), em que *O* representa o conjunto de estados associado ao hipercubo em que o servidor *i* está ocupado, P_F exibe a probabilidade de existência de fila.

$$\rho_i = \sum_{(O:o_i=1)} P_O + P_F \tag{14}$$

Outra importante medida de desempenho do sistema, a frequência total de despachos do servidor *i* no átomo *j*, inclui despachos de chamados que não aguardaram em fila, representados por $f_{ij}^{\{nq\}}$, e despachos de chamados que tiveram atrasos na fila, representados por $f_{ij}^{\{q\}}$. Considerando a probabilidade de saturação, $P_s = P_{111} + P_F$, como a probabilidade de um chamado, ao chegar no sistema, ter que aguardar em fila, e E_{ji}, o conjunto dos estados em que o servidor *i* atende um chamado no átomo *j* sem que este chamado tenha que aguardar em fila, tem-se a Equação (15).

$$f_{ij} = f_{ij}^{\{nq\}} + f_{ij}^{\{q\}} = \frac{\lambda_j}{\lambda} \sum_{j \in E_{ji}} P_O + \frac{\lambda_j}{\lambda} P_S \frac{\mu_i}{\mu} \tag{15}$$

Uma medida central de desempenho do MHF é o tempo médio de viagem. Esta medida leva em consideração complicações que influenciam no deslocamento, como por exemplo ruas de sentido único, barreiras, condições do trânsito, etc. Sendo assim, o tempo de viagem de um átomo *i* para o *j* pode ser diferente do tempo gasto para viajar do *j* para o *i*. O tempo médio de viagem do servidor *i* para o átomo *j* (t_{ij}) para chamados que não esperam na fila é apresentado na Expressão (16), onde l_{ik} é a matriz da localização dos servidores e τ_{kj} é a matriz com o tempo médio de viagem entre os átomos *k* e *j*.

$$t_{ij} = \sum_{k=1}^{N_a} l_{ik}\, \tau_{kj} \tag{16}$$

O tempo médio de viagem de uma chamada que incorre em fila pode ser calculado através da Equação (17). Nessa equação, cada um dos servidores despachados tem igual probabilidade de ser enviado para atender esse chamado (LARSON, 1974). A razão $\frac{\lambda_j}{\lambda}$ representa a probabilidade de uma chamada que incorre em algum tempo de espera em fila ser originada no átomo *j* e $\frac{\lambda_k}{\lambda}$ a probabilidade da chamada ser atendida por um servidor localizado no átomo *k*.

$$T_q = \sum_{k=1}^{N_a} \sum_{j=1}^{N_a} \frac{\lambda_k \lambda_j}{\lambda^2} \tau_{kj} \quad (17)$$

O tempo médio de viagem no sistema (*T*) inclui o tempo médio das chamadas que são prontamente atendidas, ou seja, não esperam em fila, e das chamadas que precisam esperar um determinado tempo em fila. De acordo com Takeda, Widmer e Morabito (2001) , esta medida reflete o desempenho global do sistema, visto que é independente de qual átomo foi originado o chamado e qual servidor foi enviado para realizar o atendimento. O tempo médio de viagem no sistema pode ser obtido através da Expressão (18).

$$T = \sum_{i=1}^{N} \sum_{j=1}^{N_a} f_{ij}^{nq} \, t_{ij} + P_s T_q \quad (18)$$

O tempo médio de viagem de cada servidor *i*, de acordo com Larson e Odoni (2007), pode ser estimado por meio da Equação (19). Essa expressão é uma estimativa para essa medida, uma vez que esta não pode ser determinada exatamente. Utilizando essa expressão, quanto mais a taxa de utilização do sistema se aproxima de 1, mais exato é o valor desta estimativa, sendo que a expressão retorna o valor exato quando a taxa de utilização do sistema é igual a 1 (LARSON, 1974).

$$TU_i = \frac{\sum_{j=1}^{N_a} f_{ij}^{nq} t_{ij} + \left(T_q P_S\right).\frac{\mu_i}{\mu}}{\sum_{j=1}^{N_a} f_{ij}^{nq} + \frac{\mu_i}{\mu} P_S} \quad (19)$$

2.4 Otimização com Modelo Hipercubo de Filas

Essa seção do referencial é dedicada a apresentar trabalhos da literatura que integram o MHF com procedimentos de otimização para a localização de servidores emergenciais. Boyaci e Geroliminas (2015) destacam que embora o MHF não tenha sido inicialmente desenvolvido como um modelo de otimização, ele pode ser integrado em uma metodologia, por exemplo meta-heurísticas, para atuar como um.

Batta, Dolan e Krishnamurthy (1989) relaxaram três suposições do MEXCLP: servidores operam de forma independente, servidores têm a mesma probabilidade de

ocupação e as probabilidades de ocupação são invariáveis com respeito as suas localizações. Na tentativa de relaxar a suposição de independência das probabilidades de ocupação dos servidores foi utilizado como aproximação os fatores de correção do MHF, proposto por Larson (1974), com isso, os autores formularam o AMEXCLP. Além disso, o MHF foi incorporado a um simples procedimento heurístico de substituição de vértices, para determinar um conjunto de localizações de servidores que maximiza a cobertura esperada, dando origem ao modelo de localização hipercubo (HQM). Alocar mais de um servidor em um ponto é uma possível estratégia do MEXCLP (ao contrário do MCLP).

Saydam e Aytuğ (2003) apresentaram um algoritmo genético (AG) que combina a cobertura esperada com o MHF, para resolver o MEXCLP com aumento na precisão e realismo. Em vez de utilizarem o modelo original, os autores reformularam o MEXCLP usando uma função objetivo não linear com menos variáveis.

Chiyoshi, Galvão e Morabito (2003) compararam os métodos de solução do problema de localização de máxima cobertura esperada (MEXCLP, AMEXCLP e HQM). Segundo os autores, os métodos não são estritamente comparáveis pelo fato de apresentarem diferenças estruturais nas suas funções objetivo. O MEXCLP e o AMEXCLP não levam em consideração a influência das chamadas em fila nos cálculos. Já o HQM admite influência das chamadas em fila. Os autores concluem que o AMEXCLP, que é obtido relaxando a suposição de que os servidores operam independentemente, melhora os resultados obtidos pelo MEXCLP. Por outro lado, o HQM fornece uma alternativa para lidar com chamadas em fila. O HQM é mais do que uma melhoria dos modelos MEXCLP/AMEXCLP, uma vez que fornece uma descrição mais precisa do sistema ao lidar com problemas de localização em que a contribuição das chamadas em fila é importante para o sistema.

Iannoni, Morabito e Saydam (2008) incorporaram o MHF ao AG para determinar as melhores áreas de operação para as bases do serviço médico emergencial em rodovias com sistema de *backup* parcial. Os resultados sugerem que o algoritmo AG/Hipercubo é eficaz para apoiar decisões operacionais e de projeto, por exemplo, para determinar as melhores áreas de operação para as bases do sistema das ambulâncias, minimizar o tempo médio de resposta do usuário e os desequilíbrios da carga de trabalho das mesmas. Já Iannoni, Morabito e Saydam (2009) propuseram um método que consiste em uma abordagem de duas fases, sendo que a primeira possui como objetivo localizar as bases

de ambulâncias ao longo da rodovia e, a segunda busca dimensionar as regiões de cobertura de cada base, com o intuito de minimizar o tempo médio de resposta aos usuários e/ou o desbalanceamento das cargas de trabalho das ambulâncias do sistema.

Geroliminis, Karlaftis e Skabardonis (2009) propuseram um modelo analítico (*Spatial Queueing Model* - SQM) para a localização de veículos de emergências de sistemas de transporte com alta demanda examinando os problemas de cobertura e mediana de maneira integrada e os vincula ao MHF para melhorar a eficiência e a confiabilidade do modelo. Esse modelo supera simultaneamente três limitações dos modelos existentes de fila: considera a demanda como espacial e temporária, tal como a possibilidade do servidor não estar sempre disponível quando seu serviço é requerido; considera explicitamente que as taxas de serviço não são idênticas e podem variar entre servidores, enquanto para um dado servidor pode depender da característica do incidente; relaxa a suposição que a localização do servidor e preferências de despachos são definidas a priori, isto é, é um modelo de otimização.

Geroliminis, Kepaptsoglou e Karlaftis (2011) apresentaram um modelo de localização e uma solução heurística para a implantação de unidades móveis de reparo em trânsito na cidade de Atenas, Grécia. O modelo considera a natureza estocástica de tais serviços, assumindo que uma unidade pode estar ocupada quando um incidente ocorrer. O modelo proposto integra o MHF, um modelo de localização, que visa minimizar o tempo de resposta do sistema, e um algoritmo de otimização meta-heurística (AG).

Toro-Díaz *et al*. (2013) apresentaram um modelo de otimização não linear inteira mista que integra a alocação e decisões de despacho para um SAE. O MHF é integrado à formulação matemática, considerando os elementos do sistema de fila e o fenômeno do congestionamento. Foi utilizada uma estrutura de otimização baseada AG para resolver o problema proposto. Os autores constataram que, utilizando a abordagem integrada, um pequeno benefício pode ser alcançado avaliando os dois critérios mais estudados, minimizar o tempo de resposta e maximizar a cobertura esperada.

Preocupados em fornecer um serviço de qualidade para toda a população, Toro-Díaz *et al*. (2015) propuseram um modelo buscando uniformizar o tempo médio de resposta e a carga de trabalho dos servidores. De acordo com os autores, a perspectiva da equidade inclui igualizar o tempo médio de resposta individual e a carga de trabalho dos

servidores. O modelo proposto amplia o trabalho de Toro-Díaz *et al*. (2013), solucionando os problemas de escala da metodologia de solução, visto que pode ser aplicado para estudos de caso de tamanhos reais, e introduz a possibilidade de utilização de mais de um servidor candidato por base, tornando o modelo mais realístico. O comportamento dinâmico do sistema foi modelado utilizando um procedimento de aproximação, desenvolvido por Budge, Ingolfsson e Erkut (2009), para resolver o MHF básico. Os autores utilizam a meta-heurística BT na resolução do modelo proposto.

Boyaci e Geroliminis (2015) resolveram o problema de localização de ambulâncias tratando-o como um sistema de filas espacialmente distribuídas com demanda e tempo de serviço estocásticos. Os autores expandiram o espaço de estados do MHF de Larson (1974) ao considerar 3^n estados, sendo que existem dois tipos de serviços disponíveis para cada servidor, serviço na proximidade de um servidor (área de responsabilidade) e serviço para outros locais onde o primeiro servidor responsável está ocupado durante este evento, além do estado em que o servidor está ocioso. O principal objetivo do estudo é testar a aplicabilidade do modelo Hipercubo de filas agregado (AHQM) em algoritmos de busca, visando minimizar a taxa de perda de chamados. O AHQM foi integrado nas meta-heurísticas VNS e SA.

Kim e Lee (2016) propuseram um algoritmo iterativo de otimização que incorpora o MHF, considerando cargas de trabalho não idênticas e a fração ocupacional máxima possível com base em uma suposição de operação de ambulância dependente para resolver o problema de cobertura de conjuntos de locais probabilístico (PLSCP). O objetivo do modelo proposto é minimizar o número de ambulâncias alocadas, garantindo um nível de confiabilidade determinado. As restrições de interações propostas pelos autores permitem que o algoritmo sempre encontre localizações de ambulâncias que satisfaçam a restrição de confiabilidade. Os autores também desenvolveram um algoritmo iterativo, no qual o MHF é substituído por uma simulação, a fim de obter eficiência computacional. Os experimentos computacionais mostraram que o algoritmo iterativo de otimização que incorpora o MHF possui desempenho equivalente ao algoritmo que utiliza simulação.

Akdogan, Bayindir e Iyigun (2018) utilizaram a meta-heurística AG para resolver o modelo aproximado de filas (AQM) proposto por Akdogan, Bayindir e Iyigun (2017). O modelo em questão trata-se de um problema de p-mediana com um critério de

cobertura, onde a função objetivo, minimizar o tempo médio de resposta do sistema, é calculada a partir de um sistema de filas.

Rodriguez, De La Fuente e Aguayo (2020) apresentaram o modelo de localização de facilidades e alocação de equipamentos com cobertura esperada (FLEET-EXC), que visa maximizar a cobertura da demanda esperada. Esse modelo considera distintos tipos de chamados e de servidores além de considerar políticas de despacho de acordo com as regiões. O modelo Hipercubo de Filas é utilizado para calcular a taxa de ocupação dos servidores.

Golabian e colaboradores (2021) propuseram dois modelos de localização de ambulâncias combinados ao modelo Hipercubo de Filas visando maximizar a probabilidade de cobertura. Adicionalmente, os autores utilizaram uma estratégia de otimização via simulação combinando o algoritmo *Simulated Annealing* com a simulação de eventos discretos, para resolver problemas de larga escala.

2.5 *Variable Neighborhood Search* (VNS)

O VNS é uma meta-heurística utilizada para resolver problemas de otimização combinatória cuja ideia básica é uma sistemática mudança de estruturas de vizinhança, tanto dentro de uma fase de descida para encontrar um ótimo local quanto em uma fase de perturbação para sair do vale correspondente a esse ótimo local. Originalmente projetada para encontrar soluções aproximadas de problemas de otimização combinatória, o VNS foi estendido para tratar problemas inteiros mistos, problemas não lineares e problemas não lineares inteiros (HANSEN *et al*., 2019).

O VNS foi proposto por Mladenović e Hansen (1997) para resolver um problema de roteamento de veículos com *backhauls* (VRPB). Desde então, aplicações dessa meta-heurística estão aumentando rapidamente em número e em áreas de estudo: teoria da localização (BRIMBERG *et al*., 2016), análise de cluster (COSTA; ALOISE; MLADENOVIĆ, 2017), *scheduling* (ANOKIC *et al*., 2020; FATTAHI *et al*., 2019), roteamento de veículos (ARMAS; MELIÁN-BATISTA, 2015; SARASOLA et al., 2015; SICILA *et al*., 2015), projeto de rede (MARINAKIS; MIGDALAS; SIFALERAS, 2017), dimensionamento de lotes (MENÉNDEZ *et al*., 2016; SIFALERAS; KONSTANTARAS; MLADENOVIĆ, 2015), confiabilidade (TODOSIJEVIC *et al*., 2016), etc.

É importante explicar a lógica de estruturação do VNS. Dado um problema de otimização, representado em (20).

$$min\,\{f(x)|x\ \in X, X\ \subseteq S\} \qquad (20)$$

Seja N_k (k = 1, …, k_{max}) um conjunto finito de estruturas de vizinhança pré-selecionadas e $N_k(x)$ o conjunto de soluções da *k*-ésima vizinhança de *x*. Uma solução ótima x^* (ou mínimo global) é uma solução factível onde um mínimo de (14) é alcançado, sendo $x' \in X$ um mínimo local de (14) em relação a N_k. A meta-heurística VNS é baseada em três fatos (HANSEN *et al*., 2019):

1. Um mínimo local em relação a uma estrutura de vizinhança não é necessariamente um mínimo local em relação a outra vizinhança;
2. Um mínimo global é um mínimo local em relação a todas as estruturas de vizinhanças possíveis;
3. Para muitos problemas, mínimos locais em relação a um ou vários N_k são relativamente próximos uns aos outros.

Esse último fato, que é empírico, implica que um mínimo local geralmente fornece informações sobre o global. A fim de resolver (14) utilizando diversas vizinhanças, os fatos 1-3 são usados na resolução de três diferentes tipos de problemas: determinísticos, estocásticos e ambos.

O algoritmo VNS Básico (BVNS), apresentado no Algoritmo 1, possui uma fase estocástica, representada pela seleção aleatória do ponto *x'* da *k-ésima* vizinhança da operação de agitação, isto é, $x' \in N_k(x)$. O ponto *x'* é gerado de forma aleatória no Passo 5 a fim de se evitar ciclo, o que pode ocorrer com uma regra determinística. Em seguida, o ponto *x'* passa por uma fase determinística, que é representada por uma heurística de busca local (Passo 6). Logo após essa busca local, ocorre a mudança de vizinhança, apresentada no Passo 7. Esse algoritmo é repetido *k* vezes.

Uma condição de parada é escolhida, como, por exemplo, o tempo máximo de CPU permitido (t_{max}) ou o número máximo de iterações entre duas melhorias (k_{max}). Tem sido observado que o melhor valor para o parâmetro k_{max} é geralmente 2 ou 3 (HANSEN *et al*., 2019).

Algoritmo 1 – VNS Básico

```
Função BVNS(x, k_max, t_max)
1 t ←0
2 enquanto t < t_max faça
3       k←1
4       repita
5             x' ← Shake(x, k)                          // Agitando
6             x'' ← BestImprovement(x')                 // Busca local
7             x, k ← NeighborhoodChange(x, x', k) // Muda vizinhança
8       até k = k_max
9       t ←CpuTime()
10 retorne x
```

A fim de facilitar o entendimento do BVNS, os algoritmos a seguir apresentam suas funções. O Algoritmo 2 apresenta a função *Shake*, que gera de forma aleatória um ponto x' da *k-ézima* vizinhança, isto é, $x' \in N_k(x)$.

Algoritmo 2 – Função *Shake*

```
Função Shake(x, k)
1 w ←_1 + rand(0, 1)×|N_k(x)|_
2 x' ← x^w
3 retorne x'
```

O BVNS combina mudanças de vizinhanças determinísticas e estocásticas. A parte determinística é representada por uma heurística de busca local. Essa heurística consiste em: (1) escolher uma solução inicial x; (2) encontrar uma direção de descida a partir de x (dentro da vizinhança $N(x)$); e (3) mover para o mínimo de $f(x)$ dentro de $N(x)$. Se não há direção de descida, a heurística para e, se há direção, ela é iterada. Geralmente, a *Steepest Descent*, também conhecida como *Best Improvement*, é usada. O Algoritmo 3 apresenta a função *Best Improvement*.

Algoritmo 3 – *Best improvement (steepest descent) heuristic*

```
Função BestImprovement(x)
1 repita
2       x´ ← x
3       x ← argmin y ∈ N(x´) f (y)
4 até (f (x) ≥ f (x´))
5 retorne x
```

Como o algoritmo de *Best Improvement* pode ser um processo demorado, uma alternativa é usar a heurística de *First Improvement*, apresentada no Algoritmo 4. Na heurística de *First Improvement*, os pontos $x' \in N_k(x)$ são enumerados sistematicamente e uma mudança é feita assim que uma direção de descida é encontrada.

Algoritmo 4 – *First improvement (first descent) heuristic*

```
Função FirstImprovement(x)
1 repita
2       x' ← x; i ← 0;
3       repita
4               i ← i+1
5               x ← argmin {f(x), f(x_i)}, x_i ∈ N(x)
6       até (f(x) < f(x') ou i = |N(x)|)
7 até (f(x) ≥ f(x'))
10 retorne x
```

O Algoritmo 5 examina a mudança de solução e a função de mudança de vizinhança usada na estrutura do VNS. A Função *NeighborhoodChange*() compara o valor atual ($f(x)$) com o novo valor ($f(x')$) obtido pela vizinhança k. Se uma melhoria é encontrada, o valor atual é atualizado e k é retornado para o valor inicial. Se não, a próxima vizinhança é considerada.

Algoritmo 5 – *Neighborhood Change*

```
Função NeighborhoodChange(x, x', k)
1 se  f(x') < f(x) então
2       x ← x'          // Faça um movimento
3       k ← 1           // Vizinhança inicial
4 senão
5       k ← k+1         // Próxima vizinhança
6 retorne x, k
```

No caso de instâncias muito grandes, em que a pesquisa local executada no BVNS é dispendiosa, o *Reduced* VNS (RVNS) surge como uma alternativa. No RVNS uma solução aleatória é selecionada a partir de $N_k(x)$ e não é realizada tentativa de descida a partir desta. Pelo contrário, o valor da nova solução é comparado com o do operador histórico e uma atualização ocorre no caso de melhoria. O RVNS também pode ser utilizado para encontrar soluções iniciais para problemas de grande porte antes da decomposição. O Algoritmo 6 ilustra a RVNS.

Algoritmo 6 – *Reduced* VNS (RVNS)

```
Função RVNS(x, k_max, t_max)
1 repita
2       k ← 1
3       repita
4               x' ← Shake(x, k)
5               x, k ← NeighborhoodChange (x, x', k)
6       até k = k_max
7               t ← CpuTime()
8 até t > t_max
9 retorne x
```

3 Sistema Estudado

Com o intuito de apresentar as especificidades do sistema estudado, essa seção apresenta a descrição da região estudada e da coleta de dados realizada.

3.1 Região Estudada

Ouro Preto e Mariana são duas das cidades mais antigas de Minas Gerais. De acordo com o Censo Demográfico de 2010 (IBGE, 2011), Ouro Preto possui uma população total de 70281 habitantes em uma área territorial igual a 1245,865 km^2, enquanto Mariana possui 54219 habitantes em uma área de 1194,208 km^2. Amarantina, Antônio Pereira, Cachoeira do Campo, Engenheiro Corrêa, Glaura, Lavras Novas, Miguel Burnier, Rodrigo Silva, Santa Rita de Ouro Preto, Santo Antônio do Leite, Santo Antônio do Salto e São Bartolomeu são os distritos de Ouro Preto. O município de Mariana possui os seguintes distritos: Águas Claras, Bandeirantes, Cachoeira do Brumado, Camargos, Cláudio Manoel. Furquim, Monsenhor Horta, Padre Viegas, Passagem de Mariana e Santa Rita Durão. Os distritos de Ouro Preto concentram 41,78% da população total do município e, os distritos de Mariana são responsáveis por 28,13% da população total (IBGE, 2011). A Figura 3 representa a distribuição geográfica da região estudada.

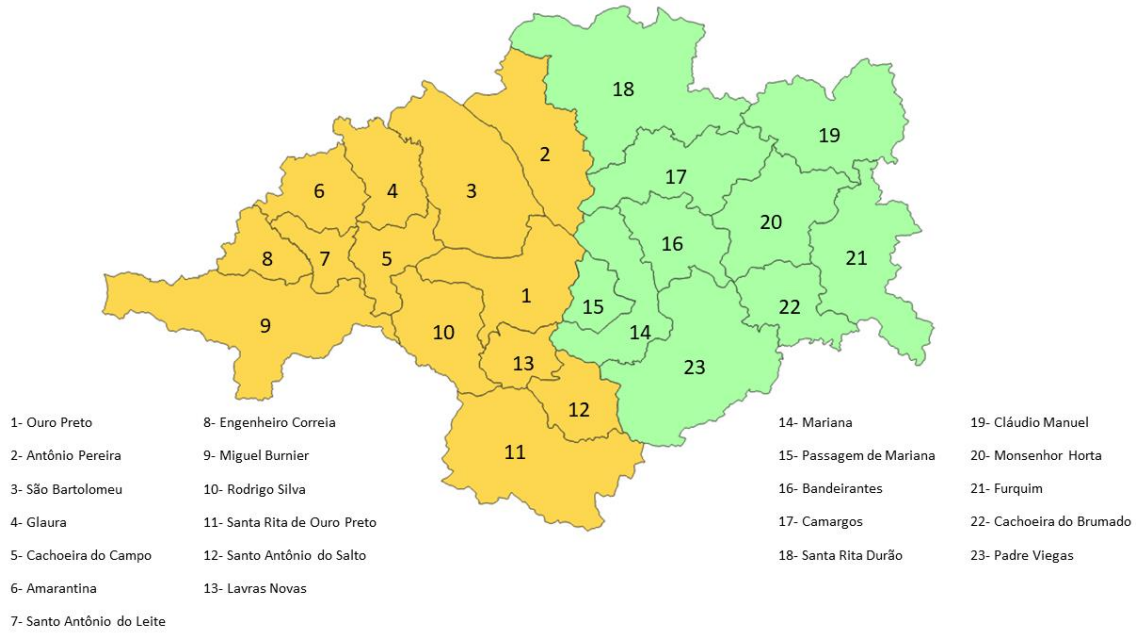

Figura 3: Distribuição geográfica da região estudada

O conjunto de ambulâncias do SAMU de Ouro Preto e Mariana é responsável pela prestação de atendimento médico emergencial em todo o território destes municípios em situações de natureza clínica, cirúrgica, traumática, obstétrica, pediátrica e psiquiátrica**.**

O município de Ouro Preto dispõe de três unidades de atendimento, sendo duas delas de suporte básico (USB) e uma delas de suporte avançado (USA), enquanto Mariana dispõe de uma unidade de atendimento básico (USB). É válido ressaltar que as USBs são compostas por no mínimo dois profissionais, sendo um condutor e um técnico ou auxiliar de enfermagem, já a USA tem a bordo um condutor, um médico e um enfermeiro.

O atual SAMU opera 24 horas por dia, 7 dias por semana e conta com 3 bases espalhadas na região, com um número fixo de ambulâncias. A primeira base está localizada na Unidade de Pronto Atendimento de Ouro Preto (UPA-OP), situada à Rua Mecânico José Português, 167, onde se encontram a USA 0009 e a USB 7555. A USB 7556 se encontra no distrito de Cachoeira do Campo, Praça Benedito Xavier, s/n, e a terceira base descentralizada é localizada na Policlínica Municipal de Mariana, Rua do Contorno s/n, onde se situa a USB 7557. Até setembro de 2017, não existia a base de Cachoeira do Campo, a USB 7556 se localizava na UPA-OP, junto com a USB 7555, e a USA 0009 situava-se no Hospital Santa Casa de Misericórdia, localizado à Rua José Moringa, 620. A USB 7557 de Mariana já se situava na Policlínica Municipal de Marina.

O atendimento do SAMU inicia-se a partir de contato telefônico gratuito da vítima ou de alguém próximo que esteja a assistindo. Neste primeiro momento são prestadas orientações básicas fornecidas por um técnico que identifica a emergência e coleta as primeiras informações sobre a vítima e sua localização. A seguir, a chamada é repassada para o médico regulador, que fornecerá instruções mais específicas e acionará as ambulâncias quando necessário.

A fim de reduzir custos, a Prefeitura de Belo Horizonte passou a ser responsável pela regulação do SAMU. Os chamados que chegam para a central reguladora são classificados pelo médico regulador em dois tipos, chamados do tipo *a*, que são os casos considerados mais graves, e chamados tipo *b*, que são casos um pouco mais simples, mas que ainda requerem uma atenção especial. Após realizar essa classificação, o médico regulador repassa o chamado para a ambulância que irá realizar o atendimento, através de uma ligação para o telefone corporativo da equipe da ambulância. É importante ressaltar que esse estudo não obteve acesso aos dados da regulação.

Por ser mais bem equipada e possuir um médico em sua tripulação, a USA é responsável por atender os chamados do tipo *a*. Caso a USA esteja ocupada, a USB disponível mais próxima do chamado é enviada para realizar o atendimento. Os chamados

do tipo *b* são atendidos preferencialmente pelas USBs, sendo que a ambulância localizada mais próxima ao chamado é enviada. Caso chegue um chamado no sistema e todas USBs estejam ocupadas, a USA não é enviada, o chamado aguarda em uma fila, esperando alguma USB finalizar seu atendimento.

Para sintetizar o processo de atendimento do SAMU de Ouro Preto e Mariana, a Figura 4 foi elaborada.

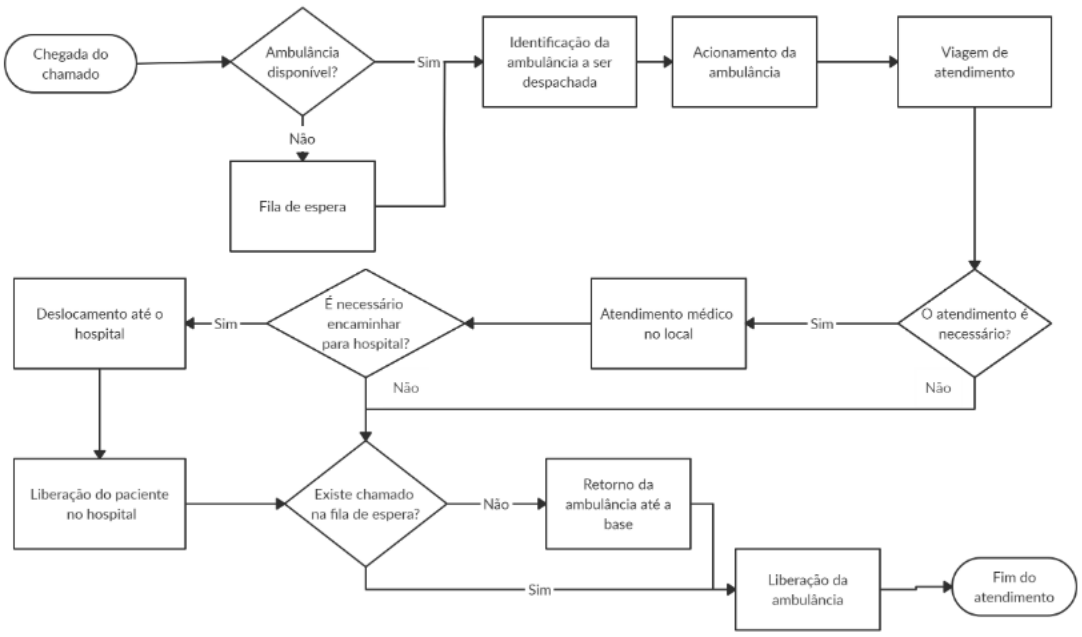

Figura 4: Processo de atendimento do SAMU

A Tabela 5 apresenta as políticas de despacho das ambulâncias. As regras são definidas por tipo de chamado e regiões geográficas. Até o final de agosto de 2017, não havia preferência entre as ambulâncias USB 7555 e USB 7556 que se localizavam na UPA-OP. Elas eram despachadas de forma aleatória. Com a mudança da USB 7556 para Cachoeira do Campo, a USB 7556 passou a ser a ambulância preferencial para os chamados do tipo *b* da região de Cachoeira do Campo e a USB 7555 a preferencial para os chamados de Ouro Preto (sede). A USB 7557, localizada em Mariana, é a ambulância preferencial para os chamados do tipo *b* de Mariana, seus distritos e de Antônio Pereira.

Depois do atendimento no local, pode ser necessário ou não remover o paciente para o hospital. Se a remoção não for necessária, caso não tenha ocorrido outra chamada no sistema, o veículo volta para a base e é liberado, já se o transporte se fizer necessário, a ambulância leva o paciente ao hospital, retorna para a base e só então é liberada.

Em relação ao processo de regionalização dos serviços do SAMU estabelecido pela Portaria nº 1010 de 21 de maio de 2012 (MINISTÉRIO DA SAÚDE, 2012), a microrregião de Ouro Preto é composta pelos municípios de Ouro Preto, Mariana, Itabirito e Diogo de Vasconcelos. Entretanto, conforme descrito anteriormente, esse processo aconteceu parcialmente apenas entre os munícipios de Ouro Preto e Mariana. O Consórcio Aliança, com sede na cidade de Belo Horizonte – Minas Gerais, é a empresa responsável por gerenciar o funcionamento do SAMU em 81 cidades mineiras. Ouro Preto e Mariana fazem parte destas cidades atendidas pelo consórcio.

Tabela 5 – Política de despacho

Cenário centralizado								
Região	Chamados avançados				Chamados básicos			
	1º	2º	3º	4º	1º	2º	3º	4º
Ouro Preto (sede)	1	2 ou 3	2 ou 3	4	2 ou 3	2 ou 3	4	-
Amarantina, Cachoeira do Campo, Engenheiro Corrêa, Glaura, Santo Antônio do Leite, São Bartolomeu	1	2 ou 3	2 ou 3	4	2 ou 3	2 ou 3	4	-
Lavras Novas, Rodrigo Silva, Santa Rita de Ouro Preto, Santo Antônio do Salto	1	2 ou 3	2 ou 3	4	2 ou 3	2 ou 3	4	-
Mariana (sede) e Passagem de Mariana	1	4	2 ou 3	2 ou 3	4	2 ou 3	2 ou 3	-
Bandeirantes, Monsenhor Horta, Padre Viegas, Cachoeira do Brumado e Furquim	1	4	2 ou 3	2 ou 3	4	2 ou 3	2 ou 3	-
Camargos, Santa Rita Durão, Águas Claras, Cláudio Manuel e Antônio Pereira	1	4	2 ou 3	2 ou 3	4	2 ou 3	2 ou 3	-
Cenário descentralizado								
Região	Chamados avançados				Chamados básicos			
	1º	2º	3º	4º	1º	2º	3º	4º
Ouro Preto (sede)	1	2	3	4	2	3	4	-
Amarantina, Cachoeira do Campo, Engenheiro Corrêa, Glaura, Santo Antônio do Leite, São Bartolomeu	1	3	2	4	3	2	4	-
Lavras Novas, Rodrigo Silva, Santa Rita de Ouro Preto, Santo Antônio do Salto	1	2	3	4	2	3	4	-
Mariana (sede) e Passagem de Mariana	1	4	2	3	4	2	3	-
Bandeirantes, Monsenhor Horta, Padre Viegas, Cachoeira do Brumado e Furquim	1	4	2	3	4	2	3	-
Camargos, Santa Rita Durão, Águas Claras, Cláudio Manuel e Antônio Pereira	1	4	2	3	4	2	3	-

1: USA 0009; 2: USB 7555; 3: USB 7556; 4: USB 7557

3.2 Coleta de Dados

A coleta e análise de dados foram autorizadas 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 na UPA-OP e na UPA-MA. Os dados coletados correspondem aos atendimentos realizados pelas quatro ambulâncias durante o ano de 2017. Esses dados foram essencialmente coletados em prontuários de atendimentos.

Esses prontuários possuem cinco campos a serem preenchidos, são eles: horário de chegada do chamado, horário de chegada ao local, horário de saída do local, horário chegada ao hospital e horário de liberação da unidade. Através das informações contidas nesses campos foi possível obter preciosas informações do serviço de atendimento, tais como o tempo de resposta do chamado, e o tempo de atendimento, representado pela diferença entre o horário de chegada ao local e o horário da saída do local.

Diante do fato do preenchimento de muitos prontuários não serem feitos de forma completa, dado que são realizados durante situações controversas de emergência, houve a necessidade de exclusão de uma parcela de dados incompletos que não permitiam a análise do comportamento do sistema. Foram definidos como chamados válidos, os chamados que apresentavam pelo menos o horário de chegada do chamado ou o horário de chegada ao local devidamente preenchido. Todos os chamados válidos que foram encaminhados ao hospital, os chamados que resultaram em óbito da vítima e os chamados em que o paciente teve alta no local foram considerados como chamados atendidos. Já os chamados em que houve encaminhamento para o hospital, compreendem os chamados válidos em que foi necessário remover a vítima para um hospital a fim de proporcionar maiores cuidados à mesma.

A Tabela 6 apresenta, para cada ambulância, o total de chamados, os chamados válidos, isto é, o número de vezes que elas foram enviadas, a porcentagem descartada de dados que não foram considerados válidos, a quantidade de atendimentos realizados, o percentual descartado dos chamados que elas foram até o local, mas não encontraram a vítima, por motivo de trote, ou da mesma ter sido transportada por terceiros, ou encontraram a vítima, mas a mesma recusou atendimento, a quantidade de atendimentos em que as vítimas foram encaminhadas para o hospital e, por fim, o percentual de descarte dos chamados atendidos mas que não foram encaminhados ao hospital pelo fato da vítima

ter vindo a óbito no local ou por não ser necessário transportar a vítima para um hospital, ou seja, apenas o atendimento em cena foi suficiente.

Tabela 6 – Dados analisados

Número de fichas analisadas	USA0009	USB7555	USB7556	USB7557	Total
Total de chamados	703	1159	983	1424	4269
Chamados válidos	609	1105	921	1345	3980
Primeiro descarte	13,37%	4,66%	6,31%	5,55%	6,77%
Atendimentos realizados	473	795	630	1072	2970
Segundo descarte	22,33%	28,05%	31,60%	20,30%	25,38%
Encaminhamento para o hospital	369	770	602	1022	2763
Percentual de atendimentos encaminhados para o hospital	78,01%	96,86%	95,56%	95,34%	93,03%

A partir da Tabela 6, é possível identificar que a USA0009 apresentou maior percentual de dados descartados em função de prontuários incompletos e esse descarte pode impactar significativamente o cálculo do intervalo de chegada dos chamados do tipo *a*. Todas as ambulâncias apresentaram uma taxa maior do que 20% no segundo descarte. Isso ocorre por diversos motivos, dentre os quais pode-se citar: a vítima é socorrida por terceiros, a vítima não foi localizada, trote e a vítima recusou atendimento. Em aproximadamente 75% dos chamados houve atendimento da vítima e estes foram considerados válidos. O fato da USA0009 apresentar um percentual de atendimentos encaminhados para o hospital consideravelmente inferior ao das USBs se deve principalmente aos pacientes falecerem antes ou durante o atendimento da equipe do SAMU. Mesmo reconhecendo a importância de tratamento adequado dos dados, um tratamento estatístico mais aprofundado de censura dos dados está fora do escopo desse trabalho.

4 Validação do *Variable Neighborhood Search*

Esta seção apresenta uma adaptação do *Variable Neighborhood Search* para tratar o problema proposto. Para isso, foram criadas instâncias testes que foram resolvidas pela meta-heurística adaptada e por um modelo matemático. Em seguida, foi feito um comparativo dos resultados obtidos pelo VNS e pelo modelo matemático, visando validar a meta-heurística proposta.

4.1 Meta-heurística Proposta

O VNS foi escolhido por se tratar de uma meta-heurística recente e por ter se mostrado mais rápida e eficiente para a resolução de diversos tipos de problemas em comparação a outras meta-heurísticas (ver, por exemplo, (GUAN; LIN, 2015; KALAYCI; KAYA, 2016; LEI, 2015; MARINAKIS; MIGDALAS; SIFALERAS, 2017; MENÉNDEZ et al., 2016; MLADENOVIĆ; TODOSIJEVIĆ; UROŠEVIĆ, 2016; SARASOLA et al., 2015; SIFALERAS; KONSTANTARAS; MLADENOVIĆ, 2015; WANG; GOLDEN; WASIL, 2019). Nos estudos citados, para distintos problemas, o VNS superou meta-heurísticas tradicionais como *Simulated Annealing*, *Busca Tabu* e Algoritmo Genético, bem como, se mostrou mais eficiente do que abordagens híbridas e outras meta-heurísticas propostas recentemente.

No presente trabalho, foi utilizada a meta-heurística VNS na sua versão básica (BVNS), apresentada no Algoritmo 1 da Seção 2.5. O BVNS foi desenvolvido utilizando a linguagem de programação Python.

Para a implementação da meta-heurística, o primeiro passo foi a geração de uma solução inicial. Para gerar uma solução inicial, foi utilizado o critério aleatório, isto é, os servidores foram alocados nas regiões aleatoriamente. Após essa alocação, é calculado o valor da cobertura corrente. Em seguida, é realizada uma agitação nessa solução inicial. A função *Shake* (Algoritmo 2) troca, de forma aleatória, a posição de um servidor, selecionado aleatoriamente. Para essa nova solução gerada, é calculado o valor da função objetivo e, caso esse valor seja superior ao valor da cobertura corrente, o valor da cobertura corrente é atualizado.

A Figura 5 apresenta um exemplo de uma região com 100 átomos e quatro servidores disponíveis. Essa figura apresenta a transição de uma solução inicial para uma solução gerada pela função *Shake*. Na solução inicial (5a), os quatro servidores estão

alocados nos átomos 1, 26, 54 e 87. Já na solução (5b), os servidores estão alocados em 1, 26, 49 e 87. Para isso, a função *Shake* selecionou aleatoriamente qual servidor seria alocado em outro átomo, nesse caso o servidor 3, e o alocou no átomo 49, escolhido aleatoriamente. Vale destacar que a meta-heurística permite que dois servidores sejam alocados em um mesmo átomo.

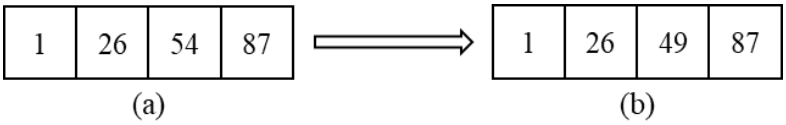

Figura 5: Transição da solução inicial para a solução gerada pela função *Shake*

No terceiro passo, a heurística de busca local troca um servidor de região por vez. Sendo assim, em cada busca local realizada, apenas um servidor sai de sua alocação e é alocado em outra região candidata, permanecendo na vizinhança da solução anterior. Esse procedimento de troca é realizado até que o algoritmo encontre o primeiro valor de cobertura da nova solução que seja maior do que o valor da cobertura corrente. Nessa fase determinística do BVNS, foi utilizada a heurística de busca local *First Improvement*, apresentada no Algoritmo 4, por ser um método de busca mais veloz do que o *Best Improvement* (Algoritmo 3).

A Figura 6 apresenta a lógica de funcionamento da heurística de busca local *First Improvement*. Na primeira solução, os servidores estavam alocados nos átomos 1, 26, 49 e 87 (Figura 6a). Após a primeira troca dentro da busca local, o servidor localizado no átomo 26 foi trocado para o átomo 95 (Figura 6b). O algoritmo comparou o valor da função objetivo da solução apresentada na Figura 6b com o valor da função da solução da Figura 6a e identificou que o valor da função objetivo da solução ilustrada na Figura 6b não era melhor do que o valor da função da Figura 6a. Dessa forma, esse procedimento de troca se repete até que o algoritmo encontre uma nova solução com valor da cobertura maior do que o valor corrente. No exemplo da Figura 6, o valor da cobertura da solução ilustrada em 6c, foi melhor dor que o valor da solução corrente (6a). Dessa forma, a heurística retorna à solução apresentada em 6c.

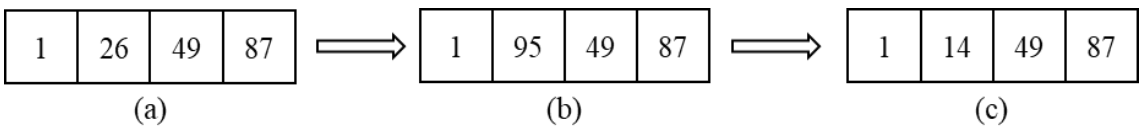

Figura 6: Procedimento de Busca Local

O passo seguinte consiste na avaliação da mudança de solução. A função de mudança de vizinhança (Algoritmo 5) compara o valor da cobertura corrente com o novo valor obtido pela vizinhança *k*. Se uma melhoria é encontrada, o valor da cobertura corrente é atualizado e *k* é retornado para o valor inicial. Se não, a próxima vizinhança é considerada. Em outras palavras, o algoritmo roda a função *Shake* para encontrar uma solução vizinha da solução anterior. No algoritmo implementado, foi definido o valor de *k* igual a 2, caso o algoritmo rode por duas vezes consecutivas a função de mudança de vizinhança e não encontre uma melhoria, o algoritmo sai do laço de iteração e o tempo de execução é verificado. Caso o tempo de execução seja maior do que o tempo máximo estipulado, a execução do algoritmo é encerrada, caso contrário, todo processo é repetido.

4.2 Modelo Matemático

Visando validar a meta-heurística proposta, foi utilizado um modelo matemático para localizar as bases de ambulâncias ao longo das regiões candidatas disponíveis e determine quais regiões de demanda cada servidor (ambulância) irá atender. Foi proposta uma alteração no modelo MALP I de ReVelle e Hogan (1989), visando contemplar as especificidades do sistema estudado no modelo matemático. Dessa forma, o modelo permite a alocação de mais de um servidor em uma única região candidata, assim como acontece no sistema estudado.

Conforme elucidado por Galvão e Morabito (2008), foi realizada uma alteração no modelo MALP I (Equações (1) a (5)), onde a Equação (5) foi substituída pelas restrições (21) e (22), que permitem que mais de um servidor seja alocado na região candidata $j \in J$ e define a natureza inteira da variável, respectivamente.

$$x_j \leq p \qquad j \in J \qquad (21)$$

$$x_j \in Z^+ \qquad j \in J \qquad (22)$$

4.3 Instâncias Utilizadas

As instâncias utilizadas foram geradas através de um algoritmo desenvolvido na linguagem de programação Python. Foram criadas 40 instâncias, agrupadas em três grupos. Para todos os grupos de instâncias, os parâmetros confiabilidade (α = 0,9) e tempo de atendimento dos servidores (t = 0,4 horas) se mantiveram constantes. Além disso, o parâmetro t_{ij} (menor distância da região j para a área de demanda i) variou, de forma aleatória seguindo a distribuição uniforme, entre 0 e 30 para as instâncias geradas. Dessa

forma, os parâmetros, apresentados na Tabela 7, sofreram alterações nos grupos das instâncias.

Tabela 7 – Parâmetros alterados entre os grupos de instâncias

Grupo 1		**Grupo 2**		**Grupo 3**	
f_i	Variou entre 0 e 1	f_i	Variou entre 0 e 1	f_i	Variou entre 0 e 0,5
S	15	S	20	S	15

Os parâmetros apresentados na Tabela 7 servem para diferenciar os grupos de instâncias. Por exemplo, no Grupo 1, o número de chamados por dia da região *i* variou entre 0 e 1 e o raio de cobertura dos servidores foi considerado igual a 15. Além dos parâmetros que foram alterados de um grupo para outro, os parâmetros regiões de demanda, regiões candidatas, número de servidores variaram dentro de um mesmo grupo de instâncias. Vale destacar que os testes foram feitos com o objetivo de abranger diferentes cenários possíveis.

As instâncias foram referenciadas por grupo (G1, G2 e G3), número de regiões de demanda e candidatas (50, ..., 2000) e número de servidores (2, …, 25). Por exemplo, a instância teste G2_150_4, trata-se de uma instância do grupo 2, com 150 regiões de demanda e candidatas a receberem 4 servidores.

4.4 Comparação entre Modelo Matemático e Meta-heurística

As instâncias apresentadas na seção 4.3 foram utilizadas no modelo matemático, apresentado na seção 4.2, e implementado no CPLEX 12.9.0.0 utilizando a linguagem AMPL, e na meta-heurística BVNS, desenvolvida utilizando a linguagem de programação Python. A solução do modelo matemático foi comparada com a solução encontrada pela meta-heurística para as instâncias teste. Os testes foram executados em um computador AMD Ryzen 7 1800X *Eight-Core Processor* 3.60GHz e 64GB de RAM, utilizando sistema operacional Windows 10 Pro.

Na Tabela 8 são apresentados os resultados obtidos e o tempo de execução para as instâncias teste geradas. Foi estabelecido um limite de 7200 segundos para a execução do modelo matemático. Caso este limite fosse ultrapassado, a execução era interrompida e a solução encontrada era selecionada para a comparação. Caso não fosse encontrada nenhuma solução dentro do tempo estabelecido, considera-se como solução não encontrada.

Para o BVNS, foram realizados 30 testes para uma mesma instância e foi estabelecido um limite de 120 segundos para a execução de cada teste. Os valores das coberturas e dos tempos computacionais da meta-heurística representam a média dos 30 testes realizados. Foram calculados os *gaps* para a cobertura considerando a meta-heurística e o modelo matemático. Ademais, foi calculada a variação percentual do tempo de execução do modelo matemático e da meta-heurística, representada pela Equação (23). Adicionalmente, foram calculados os desvios padrão da cobertura e dos tempos de execução utilizando os 30 testes das instâncias.

$$\Delta Tempo = \frac{(Tempo_{Modelo\ Matemático} - Tempo_{BVNS})}{Tempo_{Modelo\ Matemático}} * 100 \qquad (23)$$

Conforme apresentado na Tabela 8, para seis das 40 instâncias testes, o modelo matemático extrapolou o tempo limite de 7200 segundos estabelecido. Em 50% das instâncias, o BVNS encontrou um valor médio de solução com um *gap* menor que 1% em relação à solução do modelo matemático. É importante ressaltar que para algumas instâncias, vide "G1_400_8", "G1_450_9" e "G1_500_9", o modelo matemático extrapolou o tempo limite e, o BVNS encontrou, uma solução boa (*gap* < 5%) em um tempo 60 vezes menor. Em 42,50% dos testes, o *gap* encontrado ficou entre 1% a 5%. Para três instâncias, o *gap* entre a solução do BVNS e do modelo matemático foi considerado elevado (*gap* >5%), isso provavelmente se deve ao fato de o algoritmo ter sido limitado a um tempo consideravelmente pequeno.

A partir dos resultados apresentados na Tabela 8, foi possível validar a meta-heurística de localização de servidores emergenciais proposta.

Tabela 8 – Resultados computacionais

Instância	Modelo Matemático		BVNS		GAP Cobertura		Δ Tempo	
	Cobertura	Tempo	Cobertura	Tempo	Média	Desvio Padrão	%	Desvio Padrão
G1_50_2	19,27	0,05	19,27	8,24	0,00%	0,00	-16371,9%	8,33
G1_100_3	37,48	1,63	36,135	111,48	3,59%	0,80	-6739,0%	23,42
G1_150_4	65,01	12,31	60,46	>120	7,01%	0,66	-874,9%	0,01
G1_200_5	96,78	165,34	91,72	>120	5,23%	0,85	27,4%	0,02
G1_250_6	127,12	383,95	122,35	>120	3,75%	0,67	68,7%	0,03
G1_300_7	147,84	0,53	145,63	>120	1,49%	0,47	-22555,3%	0,05
G1_350_8	168,87	0,20	167,81	>120	0,62%	0,32	-59951,8%	0,07
G1_400_8	204,01	>7200	193,84	>120	4,98%	1,27	98,3%	0,09
G1_450_9	226,40	>7200	216,37	>120	4,43%	1,05	98,3%	0,11
G1_500_9	236,27	>7200	227,21	>120	3,83%	1,14	98,3%	0,11
G1_550_10	269,67	3,00	262,69	>120	2,59%	0,73	-3908,9%	0,21
G1_600_10	297,03	>7200	277,96	>120	6,42%	1,76	98,3%	0,16
G1_650_11	331,93	0,95	327,43	>120	1,35%	0,75	-12579,8%	0,18
G1_700_11	341,37	11,23	336,84	>120	1,33%	0,88	-972,3%	0,17
G1_750_12	370,06	>7200	357,24	>120	3,46%	1,75	98,3%	0,40
G1_800_12	385,57	>7200	371,01	>120	3,78%	1,84	98,3%	0,13
G1_850_13	420,84	848,94	409,94	>120	2,59%	1,61	85,8%	0,18
G1_900_13	453,75	453,67	441,48	>120	2,70%	1,73	73,4%	0,15
G1_950_14	469,72	14,92	462,70	>120	1,49%	1,42	-707,1%	0,35
G1_1000_14	492,09	7,97	484,74	>120	1,49%	1,24	-1421,4%	0,71
G1_1500_20	752,10	7,83	750,86	>120	0,16%	0,41	-1441,9%	1,03
G1_2000_25	989,82	7,20	989,04	>120	0,08%	0,48	-1635,4%	5,97
G2_50_2	22,61	0,02	22,61	7,94	0,00%	0,00	-39577,1%	6,32
G2_100_3	47,94	0,17	46,46	116,61	3,10%	0,87	-68491,9%	18,30
G2_150_4	71,86	0,8	70,71	>120	1,60%	0,33	-14901,5%	0,01
G2_200_5	99,83	0,09	99,71	95,17	0,12%	0,12	-105641,3%	39,66
G2_250_6	127,23	0,08	127,23	6,29	0,00%	0,00	-7760,0%	4,96
G2_500_9	238,94	0,27	238,94	17,88	0,00%	0,00	-6521,5%	18,64
G2_1000_14	492,09	0,089	492,09	2,48	0,00%	0,00	-2692,0%	0,02
G2_1500_20	752,10	2,88	752,10	5,83	0,00%	0,00	-102,4%	0,06
G2_2000_25	989,82	6,75	989,82	10,96	0,00%	0,00	-62,4%	0,09
G3_50_2	11,64	0,02	11,64	4,02	0,00%	0,00	-19991,1%	3,92
G3_100_3	26,52	0,24	26,32	117,76	0,74%	0,10	-48968,6%	9,24
G3_150_4	41,16	0,03	41,08	100,13	0,18%	0,07	-333651,7%	34,86
G3_200_5	49,44	0,05	49,44	17,69	0,00%	0,00	-35287,5%	19,77
G3_250_6	62,57	0,03	62,57	2,69	0,00%	0,00	-8854,9%	2,34
G3_500_9	124,50	0,19	124,16	120,20	0,27%	0,15	-63163,2%	0,12
G3_1000_14	249,57	0,59	249,57	6,44	0,00%	0,00	-992,3%	5,60
G3_1500_20	375,73	3,02	375,73	5,94	0,00%	0,00	-96,8%	0,03
G3_2000_25	489,44	4,11	489,44	11,08	0,00%	0,00	-169,7%	0,05

5 Aplicação do Modelo Hipercubo de Filas ao SAMU de Ouro Preto e Mariana

A aplicação do Modelo Hipercubo é apresentada nessa seção. Inicialmente, são analisadas as hipóteses de aplicação do modelo, em seguida, são apresentadas as equações de equilíbrio e por fim, as medidas de desempenho avaliadas.

5.1 Validação das hipóteses do Modelo Hipercubo de Filas

A aplicação do modelo Hipercubo de filas envolve a validação de suas nove hipóteses, apresentadas na Seção 2.3, considerando as características do SAMU de Ouro Preto e Mariana.

A primeira hipótese do modelo Hipercubo assume que é necessário dividir a região estudada em átomos geográficos. Essa divisão pode ser realizada de diversas maneiras, como por exemplo, considerando a divisão política de bairros e distritos, entre outros. Neste estudo, optou-se inicialmente por utilizar a divisão já existente dos distritos de Ouro Preto e Mariana. Entretanto, conforme apresentado na Figura 7, a maioria dos distritos apresentou menos de 50 chamados no ano de 2017, além do distrito de Miguel Burnier não ter apresentado nenhum chamado em função dos atendimentos serem realizados no município de Congonhas.

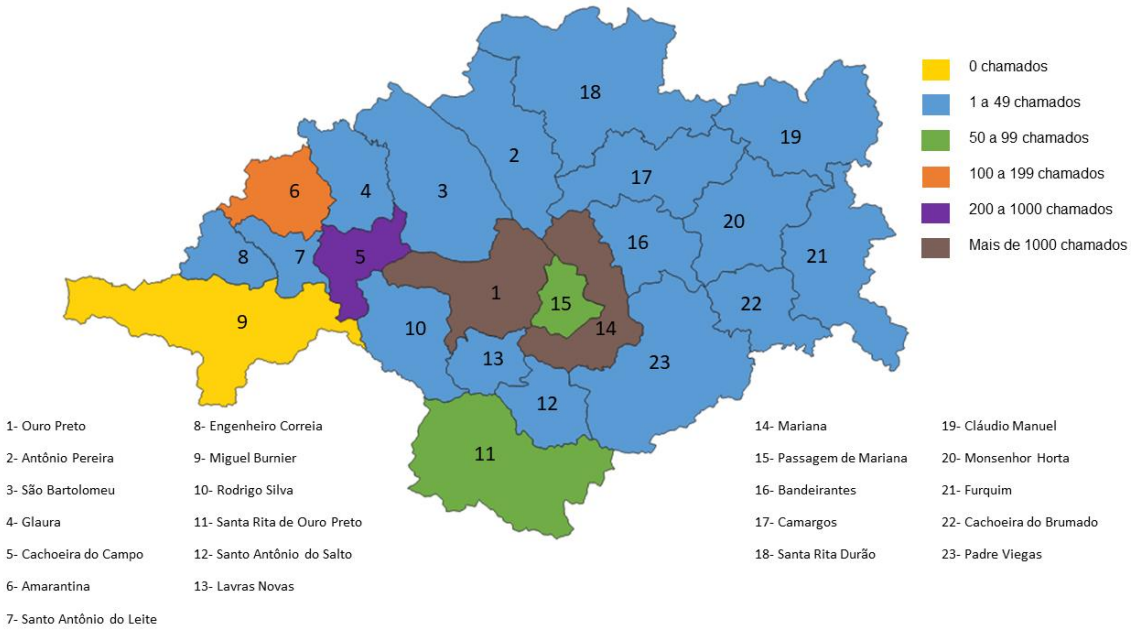

Figura 7: Mapa da localização e chamados no ano de 2017

Levando em consideração proximidade dos distritos, as estradas que ligam estes distritos e a quantidade de chamados observados durante o período de referência da coleta de dados, decidiu-se utilizar a seguinte divisão dos átomos geográficos:

1. Ouro Preto;
2. Amarantina, Cachoeira do Campo, Engenheiro Correa, Glaura, Santo Antônio do Leite, São Bartolomeu;
3. Lavras Novas, Rodrigo Silva, Santa Rita de Ouro Preto, Santo Antônio do Salto;
4. Mariana e Passagem de Mariana;
5. Bandeirantes, Monsenhor Horta, Padre Viegas, Cachoeira do Brumado e Furquim;
6. Camargos, Santa Rita Durão, Águas Claras, Cláudio Manuel e Antônio Pereira.

O distrito de Antônio Pereira, pertencente ao município de Ouro Preto, foi incluído ao átomo 7, pelo fato de se localizar mais próximo dos distritos de Mariana do que de Ouro Preto e por não apresentar uma quantidade significativa de chamados para o mesmo ser um único átomo. Miguel Burnier foi excluído da lista dos átomos, devido ao fato do distrito não ter apresentado nenhum chamado. A Figura 8 apresenta a divisão da região estudada em átomos.

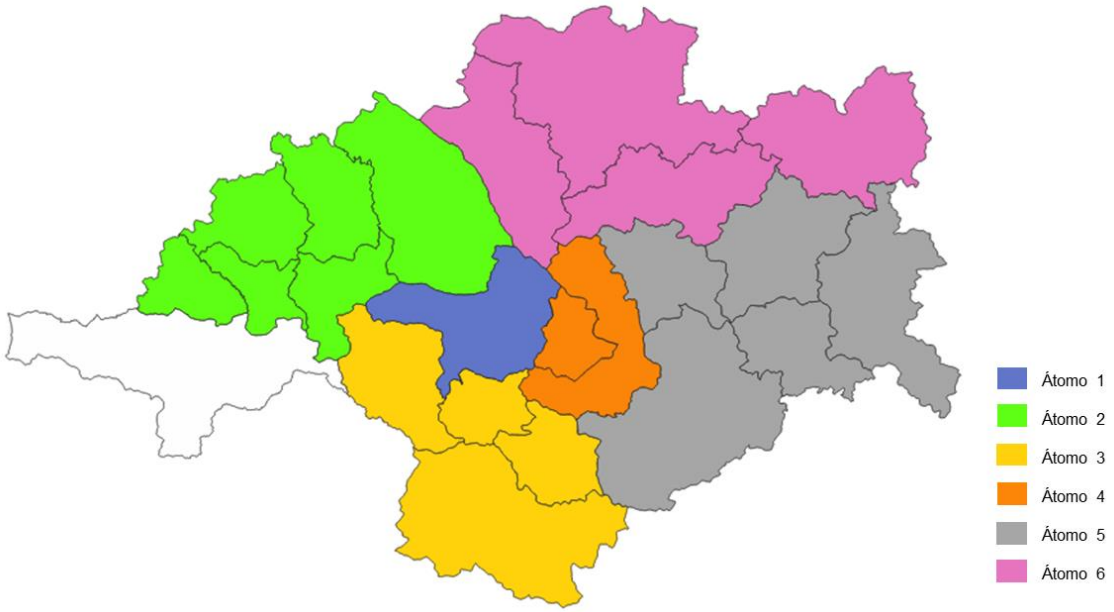

Figura 8: Divisão da região estudada em átomos geográficos

A partir dos dados coletados, foram analisados os processos de chegada dos chamados para cada átomo e de cada tipo de chamada (tipo *a* ou *b*). Neste caso, de acordo com Takeda, Widmer e Morabito (2004), o modelo Hipercubo pode ser adaptado para refletir os critérios de prioridade, bem como as regras de despacho dos servidores. Além dos

processos de chegadas dos chamados, foi possível calcular o intervalo médio entre os chamados, seu desvio padrão, o coeficiente de variação e a taxa de chegada de chamados, mostrados na Tabela 9. Nesse trabalho, assumiu-se que as taxas de chegada em cada átomo e para cada tipo de chamado seguem uma distribuição exponencial.

Tabela 9 – Dados do processo de chegada dos chamados

Átomo	Tipo de chamado	Número de observações	Intervalo médio (horas)	Desvio padrão (horas)	Coeficiente de variação	Taxa de chegada λ (chamados/hora)
1	*a*	282	31,09	35,10	1,1289	0,0322
1	*b*	1268	6,91	6,93	1,0024	0,1447
2	*a*	93	92,90	87,38	0,9406	0,0108
2	*b*	422	20,69	21,03	1,0162	0,0483
3	*a*	16	509,60	497,17	0,9756	0,0020
3	*b*	156	56,02	50,07	0,8937	0,0179
4	*a*	156	55,93	60,72	1,0856	0,0179
4	*b*	1196	7,33	7,25	0,9895	0,1364
5	*a*	7	1170,38	1521,43	1,2999	0,0009
5	*b*	104	83,41	79,42	0,9522	0,0120
6	*a*	7	1365,97	1705,39	1,2485	0,0007
6	*b*	69	126,26	130,56	1,0340	0,0079

De acordo com a terceira hipótese do MHF, os tempos médios de viagem de cada ambulância precisam ser estimados. Para auxiliar neste cálculo o aplicativo Google Maps foi utilizado. Para átomos que são formados por um conjunto de distritos, utilizou-se uma distância média entre o centroide do átomo e os demais átomos.

Sendo assim, os tempos de viagem entre os pares de átomos foram calculados de acordo com as distâncias entre os centroides de cada átomo e uma velocidade média de 60 km/h. É importante ressaltar que essa adaptação foi necessária, visto que as várias estradas que ligam os átomos às bases fixas, em sua maioria são não pavimentadas e de difícil acesso, possuem curvas sinuosas e mesmo sendo permitido ao SAMU trafegar em velocidade mais alta, ele não consegue mantê-la nessas estradas. Uma vez que as cidades de Ouro Preto e Mariana são cidades históricas com ruas de difícil acesso, possuem trânsito e semáforos, e os átomos formados por distritos apresentam considerável extensão territorial, foi considerado um tempo médio de deslocamento igual a 10 minutos em cada um desses átomos. A Tabela 10 apresenta os tempos de viagem entre os átomos, em minutos.

Tabela 10 – Tempo médio de viagem entre os átomos (min)

Átomo	1	2	3	4	5	6
1	10	21,2	29,4	14,3	31,2	55,8
2	21,2	10	47	41,2	58,1	82,6
3	29,4	47	10	45	61,9	86,4
4	14,3	41,2	45	10	22,4	41,4
5	31,2	58,1	61,9	22,4	10	63,9
6	55,8	82,6	86,4	41,4	63,9	10

O sistema em questão é composto de quatro servidores, sendo uma unidade de suporte avançado e três unidades de suporte básico, sendo que:

- O servidor 1 (USA 0009) e servidor 2 (USB 7555) estão alocados na Unidade de Pronto Atendimento de Ouro Preto (UPA-OP), situada à Rua Mecânico José Português, 167;
- O servidor 3 (USB 7556) está situado no distrito de Cachoeira do Campo, Praça Benedito Xavier, s/n.;
- A USB 7557 (servidor 4) localiza-se na Policlínica Municipal de Mariana, Rua do Contorno S/N.

Vale ressaltar que até setembro de 2017, não existia a base de Cachoeira do Campo e assim, o servidor 3 se localizava na UPA-OP junto com o servidor 2 e o servidor 1 situava-se no Hospital Santa Casa de Misericórdia, localizado à Rua José Moringa, 620. O servidor 4, por sua vez, já era alocado na Policlínica Municipal de Marina.

Para representar a localização atual dos servidores, foi utilizada a matriz l_{ik}. Os elementos dessa matriz representam a probabilidade de um servidor i quando disponível estar localizado no átomo k. Como no cenário original os servidores 1 e 2 permanecem fixos e localizados no átomo 1, o servidor 3 no átomo 2 e o servidor 4 alocado no átomo 4, tem-se inicialmente a matriz (24).

$$[l_{ik}] = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \quad (24)$$

Um dos requisitos para a aplicação do modelo Hipercubo tradicional é que o sistema envie apenas um servidor para atender ao chamado (despacho simples). No sistema estudado, o despacho simples é observado, visto que a cada chamado, apenas um servidor é enviado para realizar o atendimento. O sistema em questão admite filas, sendo

que, para efeitos de modelagem foi assumido que a fila apresenta capacidade limitada e igual a dois chamados.

As preferências de despacho dos servidores estão representadas na Tabela 11. Quando o sistema recebe um chamado do tipo *a*, o servidor 1 é enviado para realizar o atendimento. Caso ele esteja ocupado, a USB disponível mais próxima do chamado é enviada (*backup*). Já os chamados do tipo *b* são atendidos exclusivamente pelas USBs, sendo que o servidor localizado mais próximo ao chamado é enviado. Caso chegue um chamado no sistema e todas as três USBs estejam ocupadas, o servidor 1 não é enviado, o chamado aguarda em uma fila, esperando alguma USB finalizar seu atendimento. Sendo assim, o sistema funciona com *backup* parcial.

Tabela 11 – Política de despacho dos servidores

Átomo	Tipo de chamado	Primário	Secundário	Terciário	Quaternário
1	*a*	1	2	3	4
1	*b*	2	3	4	-
2	*a*	1	3	2	4
2	*b*	3	2	4	-
3	*a*	1	2	3	4
3	*b*	2	3	4	-
4	*a*	1	4	2	3
4	*b*	4	2	3	-
5	*a*	1	4	2	3
5	*b*	4	2	3	-
6	*a*	1	4	2	3
6	*b*	4	2	3	-

A hipótese VIII do MHF discorre sobre os tempos de atendimento. Esses tempos são obtidos a partir do intervalo de tempo entre o recebimento da chamada pelo servidor e o momento que o servidor termina o atendimento. É importante salientar que os dados relativos à regulação médica não foram disponibilizados para esse estudo, visto que este serviço é realizado em central de regulação independente.

Foi utilizando o teste de Tukey com nível de significância de 5% para verificar a homogeneidade dos servidores. Os resultados obtidos mostraram que os tempos de atendimento dos servidores são significativamente diferentes. O servidor 4, localizado em Mariana, não foi agrupado com as demais USBs. Esse servidor realiza a maioria de seus atendimentos na região urbana de Mariana uma vez que a população deste município se concentra nessa região (IBGE, 2011), resultando em um tempo de deslocamento menor que as USBs localizadas em Ouro Preto. O servidor 1, por sua vez, apresenta tempo de

serviço maior do que os demais que pode estar associado com as características do atendimento prestado e com o tempo de deslocamento. Esse servidor realiza atendimento em toda a região estudada e é responsável pela maioria das transferências para hospitais de referência em Belo Horizonte.

Também foi avaliado o tempo em cena dos servidores, e, os resultados sugerem que, para um nível de confiança de 95%, as USBs podem ser consideradas homogêneas, ou seja, pode-se assumir que não existe diferença estatística no tempo em cena. Já o servidor 1 apresentou tempo em cena significativamente maior ao que era esperado em função da complexidade do atendimento de chamados avançados.

A Tabela 12 apresenta o número de observações no período, o tempo médio de atendimento, o desvio padrão, o coeficiente de variação e a taxa média de atendimento para cada um dos servidores.

Tabela 12 – Dados dos tempos de atendimento

Servidor *i*	Número de observações	Tempo médio de atendimento (horas)	Desvio padrão (horas)	Coeficiente de variação	Taxa de atendimento μ (atendimentos/hora)
1	473	1,71	1,26	0,7397	0,5858
2	795	1,15	0,73	0,6362	0,8707
3	630	1,20	0,94	0,7843	0,8343
4	1072	1,01	0,74	0,7267	0,9856

Além do teste de Tukey, foram realizados testes estatísticos para verificar se as distribuições dos tempos de atendimento dos servidores poderiam ser consideradas exponenciais. Os resultados dos testes de aderência, mostraram que não se pode rejeitar a hipótese de que os tempos de atendimento dos servidores têm distribuição exponencial, com p-value de 0,05.

O modelo Hipercubo original admite que variações no tempo de atendimento devido às variações no tempo de viagem são assumidas como sendo de segunda ordem (LARSON; ODONI, 2007). Sendo assim, é necessário analisar a relação entre os tempos médios de viagem e os tempos médios de atendimento. A Tabela 13 mostra cada um desses tempos e define sua relação. Como se pode notar através da Tabela 13, a relação entre o tempo médio de viagem representa uma parcela importante do tempo de atendimento, o que sugere a necessidade de realizar o processo de calibração.

Tabela 13 – Relação entre tempo médio de serviço e tempo médio de viagem

Servidor i	Tempo médio de atendimento (horas)	Tempo médio de viagem (horas)	Relação entre tempo médio de viagem e tempo médio de atendimento
1	1,71	0,3175	0,1860
2	1,15	0,2909	0,2533
3	1,20	0,2901	0,2421
4	1,01	0,2275	0,2242

De acordo com Chiyoshi, Galvão e Morabito (2000), o processo de calibração inicia-se com o cálculo de um novo tempo médio de atendimento, o qual utiliza, ao invés do dado inicialmente utilizado para o tempo de viagem, a medida de desempenho tempo médio de viagem calculada para cada servidor. O novo tempo de atendimento é comparado com o tempo de atendimento utilizado como dado de entrada para o modelo hipercubo. Esse procedimento iterativo é repetido até que a diferença entre o tempo de atendimento calculado na iteração corrente e na iteração anterior seja menor que a tolerância estabelecida para todos os servidores. Nesse trabalho, foi utilizada a tolerância de 1% no processo de calibração dos tempos de viagem.

5.2 Equações de equilíbrio do sistema

Como no funcionamento do SAMU-OP/MA, o modelo utiliza quatro servidores distribuídos espacialmente para realizar o atendimento nos seis átomos geográficos nos quais a região estudada é dividida de acordo com uma lista fixa de preferência de despacho (Tabela 11) diferenciando os atendimentos em chamados dos tipos *a* e *b*. Esse modelo também trata explicitamente a prioridade dos chamados tipo *a* em relação ao tipo *b* na fila de espera.

Neste estudo, adotou-se a ordem da esquerda para a direita na descrição dos estados. Assim, o servidor 1 (USA 0009) assumiu a primeira posição, o servidor 2 (USB 7555) a segunda, o servidor 3 (USB 7556) a terceira e, consequentemente, o servidor 4 a quarta, ou seja, o estado {1001} indica que o primeiro e o quarto servidores estão ocupados e os servidores 2 e 3 se encontram disponíveis.

A Tabela 14 apresenta a notação utilizada no modelo proposto.

Tabela 14 – Notação utilizada no modelo

Descrição do Modelo	
n	Número de servidores (n = 4)
μ_j	Taxa de atendimento do servidor j (j = 1, ..., n)
μ	Taxa total de atendimento do sistema ($\mu = \sum_{j=1}^{n} \mu_j$)
N_a	Número de átomos (N_a= 6)
Q	Tipo de chamados (Q = a, b)
λ_{ti}	Taxa de chegada dos chamados tipo t (t = 1, …, Q) para cada átomo i (i = 1, …, N_a)
λ_i	Taxa de chegada dos chamados para cada átomo i (i = 1, …, N_a) ($\lambda_i = \sum_{t=1}^{Q} \lambda_{ti}$)
λ_t	Taxa de chegada dos chamados do tipo t (t = 1, …, Q) no sistema ($\lambda_t = \sum_{i=1}^{N_a} \lambda_{ti}$)
λ	Taxa total de chegada dos chamados no sistema ($\lambda = \sum_{i=1}^{N_a} \sum_{t=1}^{Q} \lambda_{ti}$)
P_e	Probabilidade de o sistema estar no estado e (e = {0000}, {1000}, ..., {bb1111})

Conforme citado anteriormente, em caso das três USBs se encontrarem em serviço, e o sistema receber um chamado do tipo *b*, este aguarda na fila a liberação de uma USB. Sendo assim, o sistema admite fila, mesmo quando o servidor 1 está disponível, este estado de fila é representado por {b0111}. Vale ressaltar que para efeitos de modelagem foi assumido que a fila apresenta capacidade limitada e igual a dois chamados. Essa estratégia de modelagem é justificada, uma vez que, a probabilidade de perda (P_{loss}) é muito baixa (por exemplo, no caso de quatro servidores, P_{loss} é igual 0,015%). A Figura 9 apresenta a representação do espaço de estados para o sistema em estudo.

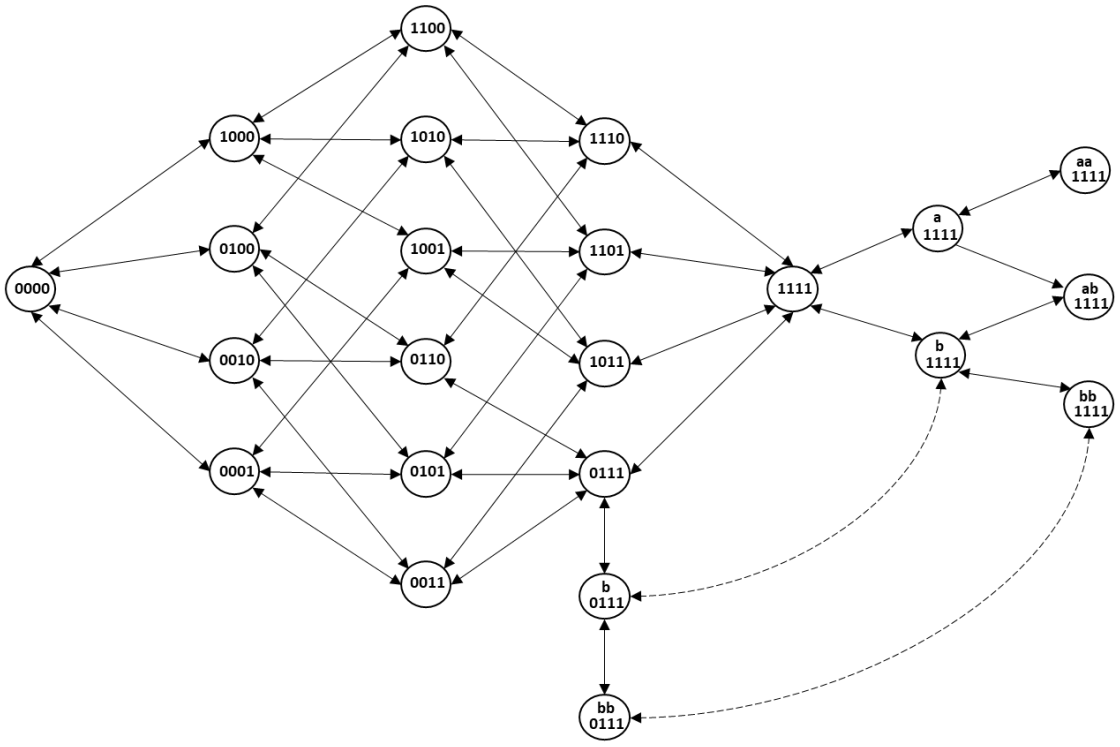

Figura 9: Representação dos estados do SAMU-OP/MA através do Modelo Hipercubo

O sistema de equações produzido utilizando o Modelo Hipercubo para este estudo de caso resulta em 23 equações lineares, uma para cada estado, cujas incógnitas são as probabilidades de equilíbrio do sistema. As equações de equilíbrio para o sistema estudado são descritas nas Equações de (25) a (47) apresentadas a seguir:

Estado {0000}

$$\lambda P_{0000} = \mu_1 P_{1000} + \mu_2 P_{0100} + \mu_3 P_{0010} + \mu_4 P_{0001} \quad (25)$$

Estado {1000}

$$(\lambda + \mu_1) P_{1000} = \lambda_a P_{0000} + \mu_2 P_{1100} + \mu_3 P_{1010} + \mu_4 P_{1001} \quad (26)$$

Estado {0100}

$$(\lambda + \mu_2) P_{0100} = (\lambda_{b1} + \lambda_{b3}) P_{0000} + \mu_1 P_{1100} + \mu_3 P_{0110} + \mu_4 P_{0101} \quad (27)$$

Estado {0010}

$$(\lambda + \mu_3) P_{0010} = \lambda_{b2} P_{0000} + \mu_1 P_{1010} + \mu_2 P_{0110} + \mu_4 P_{0011} \quad (28)$$

Estado {0001}

$$(\lambda + \mu_4) P_{0001} = (\lambda_{b4} + \lambda_{b5} + \lambda_{b6}) P_{0000} + \mu_1 P_{1001} + \mu_2 P_{0101} + \mu_3 P_{0011} \quad (29)$$

Estado {1100}

$$(\lambda + \mu_1 + \mu_2) P_{1100} = \lambda_a P_{0100} + (\lambda_1 + \lambda_3) P_{1000} + \mu_3 P_{1110} + \mu_4 P_{1101} \quad (30)$$

Estado {1010}

$$(\lambda + \mu_1 + \mu_3) P_{1010} = \lambda_a P_{0010} + \lambda_2 P_{1000} + \mu_2 P_{1110} + \mu_4 P_{1011} \quad (31)$$

Estado {1001}

$$(\lambda + \mu_1 + \mu_4) P_{1001} = \lambda_a P_{0001} + (\lambda_4 + \lambda_5 + \lambda_6) P_{1000} + \mu_2 P_{1101} + \mu_3 P_{1011} \quad (32)$$

Estado {0110}

$$\begin{aligned}(\lambda + \mu_2 + \mu_3) P_{0110} &= (\lambda_{b1} + \lambda_{b2} + \lambda_{b3}) P_{0010} + (\lambda_{b1} + \lambda_{b2} + \lambda_{b3}) P_{0100} \\ &+ \mu_1 P_{1110} + \mu_4 P_{0111}\end{aligned} \quad (33)$$

Estado {0101}

$$(\lambda + \mu_2 + \mu_4)P_{0101} = (\lambda_{b1} + \lambda_{b3} + \lambda_{b4} + \lambda_{b5} + \lambda_{b6})P_{0001} + (\lambda_{b4} + \lambda_{b5} + \lambda_{b6})P_{0100} + \mu_1 P_{1101} + \mu_3 P_{0111} \quad (34)$$

Estado {0011}

$$(\lambda + \mu_3 + \mu_4)P_{0011} = \lambda_{b2}P_{0001} + (\lambda_{b4} + \lambda_{b5} + \lambda_{b6})P_{0010} + \mu_1 P_{1011} + \mu_2 P_{0111} \quad (35)$$

Estado {1110}

$$(\lambda + \mu_1 + \mu_2 + \mu_3)P_{1110} = \lambda_a P_{0110} + (\lambda_1 + \lambda_2 + \lambda_3)P_{1010} + (\lambda_1 + \lambda_2 + \lambda_3)P_{1100} + \mu_4 P_{1111} \quad (36)$$

Estado {1101}

$$(\lambda + \mu_1 + \mu_2 + \mu_4)P_{1101} = \lambda_a P_{0101} + (\lambda_1 + \lambda_3 + \lambda_4 + \lambda_5 + \lambda_6)P_{1001} + (\lambda_4 + \lambda_5 + \lambda_6)P_{1100} + \mu_3 P_{1111} \quad (37)$$

Estado {1011}

$$(\lambda + \mu_1 + \mu_3 + \mu_4)P_{1011} = \lambda_a P_{0011} + \lambda_2 P_{1001} + (\lambda_4 + \lambda_5 + \lambda_6)P_{1010} + \mu_2 P_{1111} \quad (38)$$

Estado {0111}

$$(\lambda + \mu_2 + \mu_3 + \mu_4)P_{0111} = \lambda_b P_{0011} + \lambda_b P_{0101} + \lambda_b P_{0110} + \mu_1 P_{1111} + (\mu_2 + \mu_3 + \mu_4)\, P_{b0111} \quad (39)$$

Estado {1111}

$$(\lambda + \mu)P_{1111} = \lambda_a P_{0111} + \lambda P_{1011} + \lambda P_{1101} + \lambda P_{1110} + (\mu_2 + \mu_3 + \mu_4)P_{b1111} + \mu P_{a1111} \quad (40)$$

Cauda {0111}

Estado {*b*0111}

$$(\lambda + \mu_2 + \mu_3 + \mu_4)P_{b0111} = \lambda_b P_{0111} + \mu_1 P_{b1111} + (\mu_2 + \mu_3 + \mu_4)P_{bb0111} \quad (41)$$

Estado {*bb*0111}

$$(\lambda_a + \mu_2 + \mu_3 + \mu_4)P_{bb0111} = \lambda_b P_{b0111} + \mu_1 P_{bb1111} \quad (42)$$

Cauda {1111}

Estado {*a*1111}

$$(\lambda + \mu)P_{a1111} = \lambda_a P_{1111} + \mu P_{aa1111} \quad (43)$$

Estado {*b*1111}

$$(\lambda + \mu)P_{b1111} = \lambda_b P_{1111} + \lambda_a P_{b0111} + \mu P_{ab1111} + (\mu_2 + \mu_3 + \mu_4)P_{bb1111} \quad (44)$$

Estado {*aa*1111}

$$\mu P_{aa1111} = \lambda_a P_{a1111} \quad (45)$$

Estado {*ab*1111}

$$\mu P_{ab1111} = \lambda_b P_{a1111} + \lambda_a P_{b1111} \quad (46)$$

Estado {*bb*1111}

$$\mu P_{bb1111} = \lambda_b P_{b1111} + \lambda_a P_{bb0111} \quad (47)$$

Como esse sistema de equações é indeterminado, é necessário fazer a substituição de uma das equações pela equação de normalização, como mostra a Expressão (48).

$$\begin{aligned} P_{0000} + P_{1000} + P_{0100} + P_{0010} + P_{0001} + P_{1100} + P_{1010} + P_{1001} + P_{0110} + P_{0101} \\ + P_{0011} + P_{1110} + P_{1101} + P_{1011} + P_{0111} + P_{1111} + P_{b0111} \\ + P_{bb0111} + P_{a1111} + P_{b1111} + P_{aa1111} + P_{ab1111} + P_{bb111} = 1 \end{aligned} \quad (48)$$

Para a resolução do sistema de equações do MHF, foi utilizado o método de Gauss-Seidel, por se tratar de um método iterativo de rápida convergência para resolução de sistemas de equações lineares, para obter as probabilidades de estado associadas ao MHF.

5.3 Medidas de desempenho

Conforme mencionado anteriormente, a partir das probabilidades dos estados, podem ser calculadas diversas medidas de desempenho para o sistema analisado. Entretanto, as medidas frequência de despachos do servidor *i* no átomo *j*, tempo médio de viagem de uma chamada que incorre em fila, tempo médio de viagem dos servidores e o tempo médio de viagem no sistema, descritas na Seção 2.3.1, precisam ser adaptadas para serem utilizadas nesse sistema. Essas adaptações são necessárias para tratar características específicas do sistema estudado (*backup* parcial) e estratégia de modelagem (número de clientes na fila).

Além dessas medidas, a probabilidade de perda também deve ser calculada ao considerar que o sistema utiliza *backup* parcial e limita a capacidade da fila de espera a dois usuários. A Equação (49) apresenta o cálculo da probabilidade de perda.

$$P_{loss} = \frac{\lambda_b}{\lambda} * P_{bb0111} + P_{aa1111} + P_{ab1111} + P_{bb1111} \quad (49)$$

As probabilidades dos estados também são utilizadas no cálculo das frequências de despachos do servidor *i* ao átomo *j* para atender chamados que não incorrem em fila de espera. As Equações (50) e (51) apresentam o cálculo dessas frequências de despachos considerando a probabilidade de perda para chamados *a* e *b*, respectivamente. O conjunto E_{ij} representa o conjunto de estados nos quais o servidor *i* é o primeiro servidor disponível na lista de preferência de despacho para o átomo *j*.

$$f_{ija}^{nq} = \frac{\frac{\lambda_{ja}}{\lambda}\sum_{B\in E_{ij}} P_B}{1 - P_{loss}} \quad (50)$$

$$f_{ijb}^{nq} = \frac{\frac{\lambda_{jb}}{\lambda}\sum_{B\in E_{ij}} P_B}{1 - P_{loss}} \quad (51)$$

As frequências de despachos do servidor *i* no átomo *j* de chamados que incorrem em fila de espera para os tipos de chamado *a* e *b* podem ser encontradas a partir das Equações (52) e (53). O conjunto E_{ij}^2 representa o conjunto de estados nos quais o servidor *i* é o primeiro servidor a ser enviado ao átomo *j* para atender um chamado que incide em fila de espera. O parâmetro μ_j^s corresponde à soma da taxa de atendimento dos servidores que podem atender no átomo *j*.

$$f_{ija}^{q} = \frac{\frac{\lambda_{ja}}{\lambda}\frac{\mu_i}{\mu_j^s}\sum_{B\in E_{ij}^2} P_B}{1 - P_{loss}} \quad (52)$$

$$f_{ijb}^{q} = \frac{\frac{\lambda_{jb}}{\lambda}\frac{\mu_i}{\mu_j^s}\sum_{B\in E_{ij}^2} P_B}{1 - P_{loss}} \quad (53)$$

A Equação (54) apresenta a fórmula utilizada para calcular o tempo de viagem para chamados que aguardam em fila em sistemas com *backup* parcial. O conjunto I_i representa todos os átomos que o servidor *i* atende.

$$t_{ij}^{q} = \sum_{k=1}^{N_a} \frac{\lambda_k}{\sum_{t\in I_i}\lambda_t}\tau_{kj} \quad (54)$$

O tempo médio de viagem de cada servidor *i* (TU_i), conforme apresentado na Equação (55), é calculado a partir de uma aproximação.

$$TU_i = \frac{\sum_{j=1}^{N_a} f_{ij}^{nq} t_{ij}^{nq} + \sum_{j=1}^{N_a} f_{ij}^{q} t_{ij}^{q}}{\sum_{j=1}^{N_a} f_{ij}^{nq} + \sum_{j=1}^{N_a} f_{ij}^{q}} \quad (55)$$

O tempo médio de viagem no sistema pode ser obtido a partir da Equação (56).

$$T = \sum_{i=1}^{N} \sum_{j=1}^{N_a} f_{ij}^{nq} t_{ij}^{nq} + \sum_{i=1}^{N} \sum_{j=1}^{N_a} f_{ij}^{q} t_{ij}^{q} \quad (56)$$

5.4 Resultados da aplicação do Modelo Hipercubo de Filas ao SAMU de Ouro Preto e Mariana

Nesta seção são apresentados os resultados da aplicação do MHF ao SAMU-OP/MA.

5.4.1 Comparação das configurações atual e anterior do sistema

Conforme descrito na Seção 3.1, na configuração anterior, os servidores 2 e 3 se localizavam no átomo 1 e não havia preferência de despacho entre eles, ou seja, esses servidores eram enviados de forma aleatória. Na configuração atual, o servidor 3 foi transferido para o átomo 2 e, essa mudança, levou também a uma mudança na política de despacho de servidores. A Tabela 11, apresentada na Seção 5.1, ilustra a política atual de despacho dos servidores e a Tabela 15 apresenta a política de despacho dos servidores da configuração anterior.

Tabela 15 – Política de despacho dos servidores na configuração anterior

Átomo	Tipo de chamado	Primário	Secundário	Terciário	Quaternário
1	*a*	1	2 ou 3	2 ou 3	4
1	*b*	2 ou 3	2 ou 3	4	-
2	*a*	1	2 ou 3	2 ou 3	4
2	*b*	2 ou 3	2 ou 3	4	-
3	*a*	1	2 ou 3	2 ou 3	4
3	*b*	2 ou 3	2 ou 3	4	-
4	*a*	1	4	2 ou 3	2 ou 3
4	*b*	4	2 ou 3	2 ou 3	-
5	*a*	1	4	2 ou 3	2 ou 3
5	*b*	4	2 ou 3	2 ou 3	-
6	*a*	1	4	2 ou 3	2 ou 3
6	*b*	4	2 ou 3	2 ou 3	-

Os resultados, após a calibração, para os tempos médio de viagem dos servidores e no sistema são apresentados na Tabela 16. Já a Figura 10 apresenta um comparativo entre as cargas de trabalho dos servidores nas configurações avaliadas. Esses resultados

refletem a localização dos servidores e as políticas de despacho adotadas em cada configuração.

Tabela 16 – Tempos médio de viagem nas configurações atual e anterior do SAMU

	Tempo médio de viagem no sistema	**Tempo médio de viagem do servidor (TU_i)**			
		1	**2**	**3**	**4**
Configuração atual	13,61	14,46	12,89	16,42	12,86
Configuração anterior	13,88	14,46	14,51	14,59	12,86

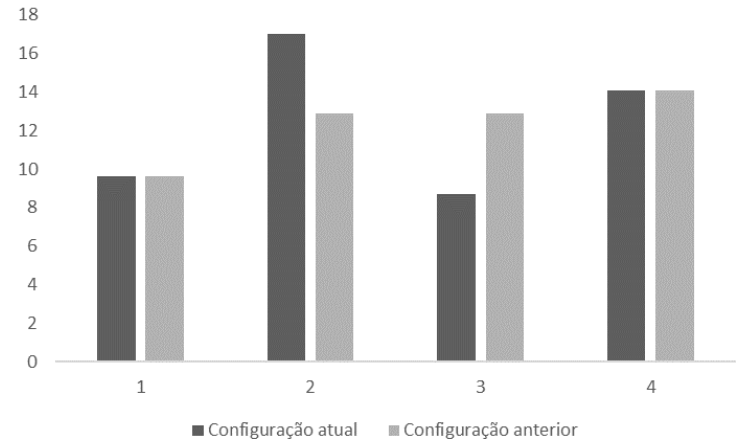

Figura 10: Taxas de ocupação dos servidores nas configurações atual e anterior

Foi observado que a configuração atual de funcionamento do SAMU-OP/MA resultou em um menor tempo médio de viagem no sistema. Já os tempos médio de viagem dos servidores mostram que a mudança de localização e política de despacho dos servidores na configuração atual influenciou no aumento do tempo de viagem do servidor 3 e diminuição do servidor 2. Entretanto, conforme apresentado na Figura 10, na configuração atual destaca-se o desbalanceamento entre os servidores 2 e 3, sendo que, o servidor 2 se encontra ocupado 17,03% do tempo e o servidor 3 apresenta uma taxa de ocupação igual a 8,71%.

5.4.2 Comparação da política atual de despachos com a política de despachos baseada na menor distância

Uma das hipóteses de aplicação do Modelo Hipercubo de Filas é a utilização de uma lista fixa de preferências de despacho. Nesse cenário, foi avaliada a política de despacho baseada na menor distância, que utiliza uma lista fixa de preferências de despacho construída baseada na distância entre os átomos ao invés de utilizar a lista de despachos baseada no funcionamento atual do sistema (Tabela 11).

A Tabela 17 apresenta os resultados do tempo médio de viagem dos servidores e tempo médio de viagem do sistema para a configuração atual do SAMU-OP/MA considerando a política atual de despachos e a política de despachos que considera a menor distância. A política de despachos baseada na menor distância resulta em um tempo médio de viagem no sistema inferior à política de despachos utilizada atualmente no sistema.

Tabela 17 – Tempos médio de viagem utilizando as políticas de despacho atual e baseada na menor distância

	Tempo médio de viagem no sistema	**Tempo médio de viagem do servidor (TU_i)**			
		1	**2**	**3**	**4**
Configuração atual	13,61	14,46	12,89	16,42	12,86
Configuração política menor distância	13,43	14,46	12,90	14,70	13,33

Esse resultado era esperado, uma vez que na política de despacho utilizada atualmente, em caso de incorrência de um chamado do tipo *b* proveniente do átomo 1 quando o servidor 2 se encontra ocupado e os servidores 3 e 4 disponíveis, o servidor 3 se desloca do átomo 2 ao átomo 1 para realizar o atendimento (com tempo de deslocamento igual a 21,2 minutos). Nessa situação, ao utilizar a política de despacho baseada na menor distância, o servidor 4 se desloca do átomo 4 ao átomo 1 para prestar o serviço (com tempo de deslocamento igual a 14,3 minutos).

Assim como no cenário anterior, nesse cenário foi traçada uma comparação entre as taxas de ocupação dos servidores utilizando a política de despacho atual e a política de despacho baseada na menor distância. Conforme ilustrado na Figura 11, a configuração com a política de despachos baseada na menor distância influenciou especialmente na carga de trabalho dos servidores 3 e 4. Vale destacar que foi observado aumento na taxa de ocupação do servidor 4 e diminuição da taxa de ocupação do servidor 3.

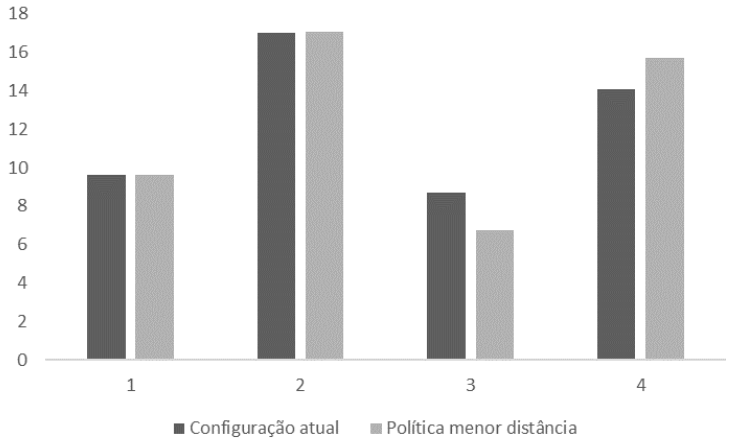

Figura 11: Taxas de ocupação dos servidores utilizando as políticas de despacho atual e baseada na menor distância

6 Abordagem combinando o VNS com o Modelo Hipercubo de Filas

Esta seção apresenta a abordagem construída, combinando o *Variable Neighborhood Search* com o modelo Hipercubo de Filas. Inicialmente, são apesentados os detalhes da incorporação do MHF ao VNS e, posteriormente, são apresentados os resultados alcançados a partir da abordagem desenvolvida.

6.1 Incorporação do MHF ao VNS

O modelo Hipercubo de Filas foi incorporado à meta-heurística VNS para otimizar a localização das ambulâncias no sistema emergencial de saúde estudado considerando as particularidades da política de despacho de servidores.

Existem estudos na literatura que combinam meta-heurísticas com o modelo Hipercubo de Filas (ver Seção 2.4), especialmente considerando o problema de localização de servidores. A Tabela 18 apresenta os trabalhos que combinam meta-heurísticas com o MHF, seus objetivos e os métodos de otimização utilizados.

Tabela 18 – Estudos que combinam meta-heurísticas com o MHF

Autores	**Objetivo**	**Método de otimização**
Batta, Dolan e Krishnamurthy (1989)	Maximizar cobertura esperada	Procedimento heurístico de substituição de vértices
Saydam e Aytuğ (2003)	Maximizar cobertura esperada	Algoritmo Genético
Iannoni, Morabito e Saydam (2008)	Minimizar tempo médio de viagem do sistema Minimizar frequência de chamados que levam mais de 10 minutos para serem atendidos Minimizar taxa de ocupação dos servidores	Algoritmo Genético
Iannoni, Morabito e Saydam (2009)	Minimizar tempo médio de viagem do sistema Minimizar taxa de ocupação dos servidores	Algoritmo Genético
Geroliminis, Karlaftis e Skabardonis (2009)	Minimizar tempo médio de viagem do sistema	Algoritmo Genético
Geroliminis, Kepaptsoglou e Karlaftis (2011)	Minimizar tempo médio de viagem do sistema	Algoritmo Genético
Toro-Díaz *et al*. (2013)	Minimizar tempo médio de viagem do sistema	Algoritmo Genético
Toro-Díaz *et al*. (2015)	Minimizar tempo médio de viagem do sistema Minimizar desigualdade entre os servidores (Índice de Gini)	Busca Tabu
Boyaci e Geroliminis (2015)	Minimizar percentual de serviços interdistritos	VNS e *Simmulated Annealing*
Kim e Lee (2016)	Minimizar número de ambulâncias	Algoritmo Iterativo de Otimização
Akdogan, Bayindir e Iyigun (2018)	Minimizar tempo médio de viagem do sistema	Algoritmo Genético
Rodriguez, De La Fuente e Aguayo (2020)	Maximizar cobertura da demanda esperada	Algoritmo Iterativo de Otimização
Golabian *et al.* (2021)	Maximizar cobertura da demanda esperada	*Simmulated Annealing*

Adicionalmente, ao integrar o MHF ao método de otimização é necessário escolher o objetivo a ser utilizado para avaliar as soluções geradas. A Tabela 18 apresenta um resumo dos objetivos utilizados em trabalhos similares. É possível perceber que os objetivos mais comuns são minimizar o tempo médio de viagem do sistema e maximizar a cobertura esperada. Além disso, Zaffar *et al.* (2016) afirmam que os trabalhos mais recentes na otimização de sistemas médicos emergenciais focam na otimização da sobrevivência do paciente.

Nesse trabalho, a abordagem foi desenvolvida com o objetivo de minimizar o tempo médio de viagem do sistema utilizando a Equação (55). O VNS tem se mostrado um método de sucesso para encontrar a solução de problemas de localização em grande escala. O algoritmo proposto foi baseado no Algoritmo 1 da Seção 2.5.A Figura 12 apresenta os passos do algoritmo construído, incorporando o MHF ao VNS.

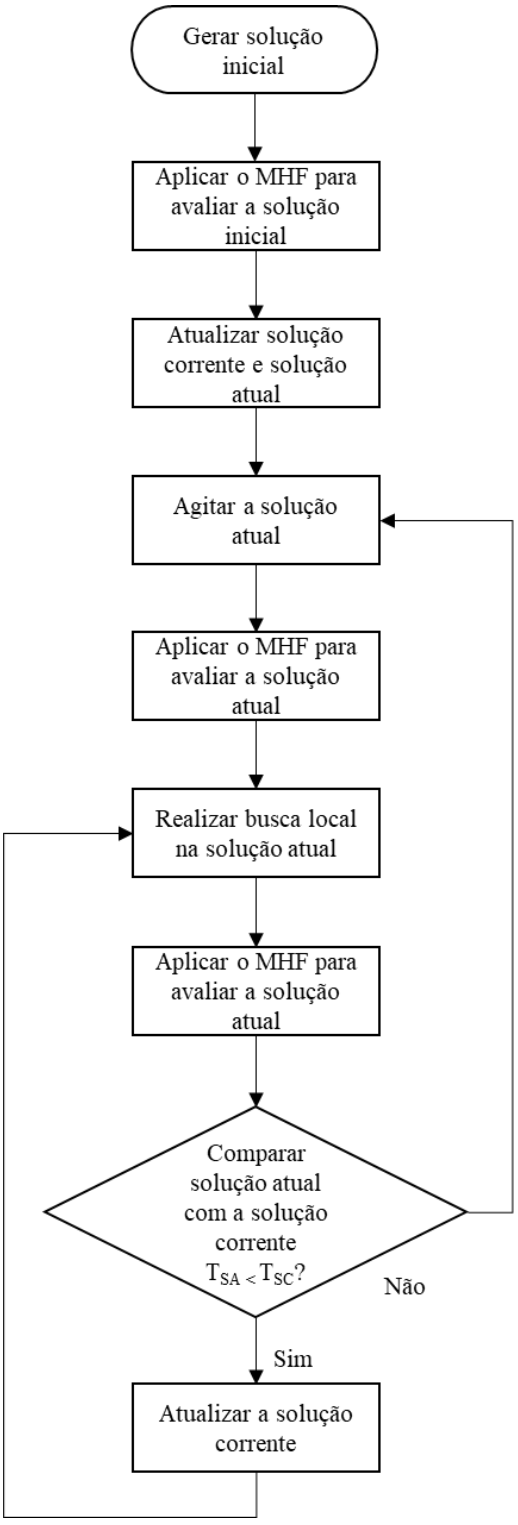

Figura 12: Algoritmo VNS/MHF

Conforme apresentado na Figura 12, o primeiro passo do algoritmo está relacionado à geração da solução inicial. Nesse passo, são definidos os parâmetros de entrada do algoritmo: a quantidade de átomos geográficos, o número de servidores do sistema, as taxas de chegada dos chamados do tipo *a* e *b*, a taxa de atendimento dos servidores e a matriz que contempla os tempos médios de viagem entre os átomos (τ_{kj}).

No algoritmo proposto, a alocação inicial dos servidores nos átomos geográficos (matriz l_{ik}) é feita de maneira aleatória, sendo que, é permitido que um átomo comporte mais de um servidor. Ao final da geração da solução inicial, o algoritmo retorna a matriz l_{ik}.

Após gerar a solução inicial, o algoritmo avança para o passo referente à avalição da solução inicial a partir do MHF. Nesse passo, o algoritmo utiliza a matriz l_{ik} e a matriz τ_{kj} para definir a política de despacho dos servidores. Na lógica utilizada, após a multiplicação das matrizes, o servidor mais próximo do átomo avaliado é considerado prioritário na lista de preferências desse átomo. Ou seja, essa abordagem utiliza uma lista dinâmica ao invés da lista fixa de preferência de despacho. No entanto, é importante ressaltar que conforme mencionado anteriormente, as USAs são os servidores preferenciais para realizar o atendimento dos chamados do tipo *a* e elas não são enviadas para atender chamados do tipo *b* (*backup* parcial). Sendo assim, essas particularidades foram consideradas na definição da política de despacho dos servidores.

Definida a política de despacho dos servidores, o algoritmo preenche a matriz do sistema de equações do MHF apresentado na Seção 5.2 para a configuração analisada. Em seguida, a última linha da matriz é substituída pela equação de normalização (47), para que o sistema passe a ser possível e determinado. O sistema resultante é então resolvido utilizando o método de Gauss-Seidel resultando em um vetor de soluções com as probabilidades dos estados do sistema. A partir dessas probabilidades, o tempo médio de viagem do sistema é calculado para a configuração analisada.

Conforme mencionado anteriormente, ao analisar a relação entre os tempos médios de viagem e atendimento no sistema analisado, ver Tabela 13, observou-se que o tempo de viagem representa uma parcela importante do tempo de atendimento nesse sistema. Em função disso, foi aplicado o processo de calibração dos tempos de viagem (CHIYOSHI; GALVÃO; MORABITO, 2000; RODRIGUES, 2014) utilizando uma tolerância de 1%.

Concluída a avaliação da solução inicial a partir do MHF, a solução corrente e a solução atual são atualizadas, considerando a alocação inicial dos servidores e o tempo médio de viagem do sistema identificado durante a aplicação do MHF.

Em seguida à atualização da solução corrente e solução atual, o algoritmo entra nos passos de otimização propriamente ditos. Seguindo a lógica do VNS, apresentada na

Seção 2.5, a solução atual é agitada utilizando a Função *Shake* (Algoritmo 2), ou seja, um servidor é selecionado aleatoriamente e transferido para outra localização também selecionada aleatoriamente. O MHF é aplicado na solução agitada (solução atual) e o tempo médio de viagem do sistema para essa nova configuração de alocação dos servidores é encontrado.

No próximo passo, foi utilizada a heurística de busca local *First Improvement*, (Algoritmo 4) para realizar a troca de um servidor de átomo por vez. Dessa forma, a cada busca local realizada, apenas um servidor sai de sua alocação e é alocado em outro átomo. Após a alteração da alocação do servidor, o MHF é utilizado para avaliar o tempo médio de viagem dessa nova configuração. Caso o tempo médio de viagem da solução atual seja inferior ao tempo médio de viagem da solução agitada, o procedimento de troca é encerrado, caso contrário, ele é realizado até que o algoritmo encontre o primeiro valor do tempo médio de viagem da nova solução que seja menor do que o valor do tempo médio de viagem da solução agitada.

Ao final da heurística de busca local, a solução atual (T_{SA}) é comparada à solução corrente (T_{SC}), utilizando a função *Neighborhood Change* (Algoritmo 5). Caso o tempo médio de viagem da solução atual seja inferior ao tempo médio de viagem da solução corrente, a solução corrente é atualizada, *k* é retornado para o valor inicial e o algoritmo retorna à heurística de busca local. Caso contrário, a próxima vizinhança é considerada, com isso, o algoritmo retorna ao passo da agitação da solução atual.

O algoritmo é executado até o tempo limite determinado (t_{max}). Ao final da execução, o algoritmo retorna o menor tempo médio de viagem do sistema e a alocação dos servidores referente à melhor solução identificada.

6.2 Implementação e resultados do Algoritmo VNS/MHF

O Algoritmo VNS/MHF, descrito na Figura 12 e detalhado na Seção 6.1, foi implementado na linguagem de programação Python. Para a execução do algoritmo, foi definido como critério de parada um tempo máximo igual a 300 segundos. Os testes foram executados em um computador AMD Ryzen 7 1800X *Eight-Core Processor* 3.60GHz e 64GB de RAM, utilizando sistema operacional Windows 10 Pro.

Os dados do SAMU-OP/MA foram utilizados no Algoritmo VNS/MHF. As taxas de chegada dos chamados dos tipos *a* e *b* (chamados/hora) dos seis átomos do sistema

seguem os valores apresentados na Tabela 9. Os tempos de deslocamento entre os átomos, matriz τ_{kj}, estão apresentados na Tabela 10. E, as taxas de atendimento dos quatros servidores (atendimentos/hora) incluídas no algoritmo estão apresentadas na Tabela 12.

Em seguida, são apresentados os resultados da implementação do Algoritmo VNS/MHF e os cenários alternativos avaliados utilizando abordagem construída.

6.2.1 Análise do SAMU-OP/MA utilizando o Algoritmo VNS/MHF

Na Figura 13 são apresentados os átomos em que os servidores foram alocados após a execução do Algoritmo VNS/MHF, ou seja, a solução final do algoritmo. Vale destacar que a solução final encontrada pelo algoritmo de otimização, segue a mesma distribuição da configuração atual do SAMU-OP/MA, com a diferença que na solução do algoritmo, o servidor 4 foi alocado no átomo 1 e o servidor 2 no átomo 4, já na configuração atual, o servidor 4 foi alocado no átomo 4 e o servidor 2 no átomo 1.

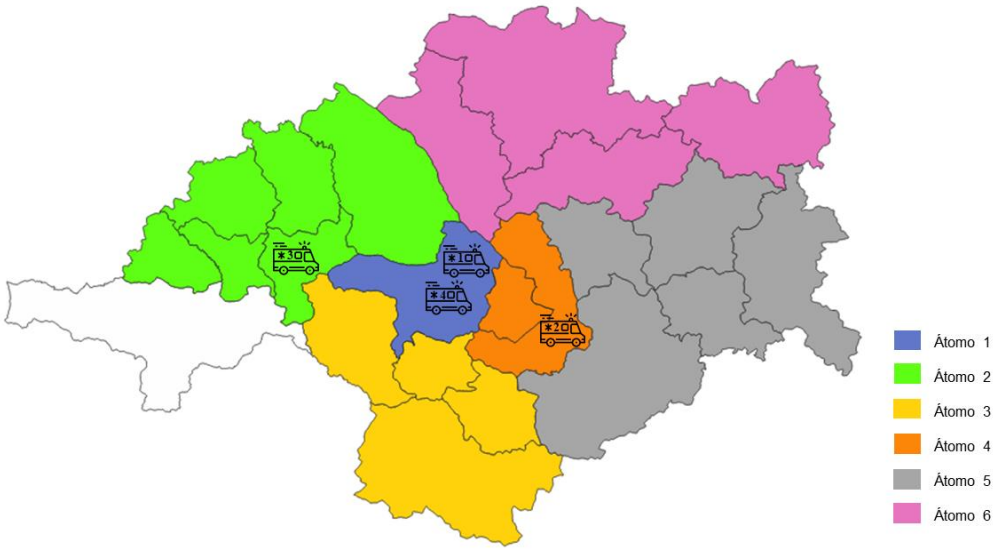

Figura 13: Localização dos servidores na solução final

A Figura 14 apresenta as taxas de ocupação obtidas por meio do Algoritmo VNS/MHF e as taxas de ocupação da configuração atual do SAMU-OP/MA. Conforme apresentado na Figura 14, houve uma redução significativa de 22,83% na taxa de ocupação do servidor 3 e um aumento de 14,80% na taxa de ocupação do servidor 4. Os servidores 1 e 2 apresentaram taxas de ocupação próximas às taxas da configuração atual.

Conforme mencionado na Seção 6.1, optou-se por utilizar uma lista dinâmica de preferência de despachos, que considera o despacho do servidor mais próximo do átomo avaliado. Ao utilizar essa lista dinâmica, o servidor 3 passou a ser despachado para atender um menor número de chamados, especialmente chamados mais distante da sua base, enquanto o servidor 4 passou a atender um maior número de chamados, inclusive na sede do município de Ouro Preto (átomo 1) quando o servidor 2 está ocupado, chamados tradicionalmente atendidos pelo servidor 3 (Tabela 11).

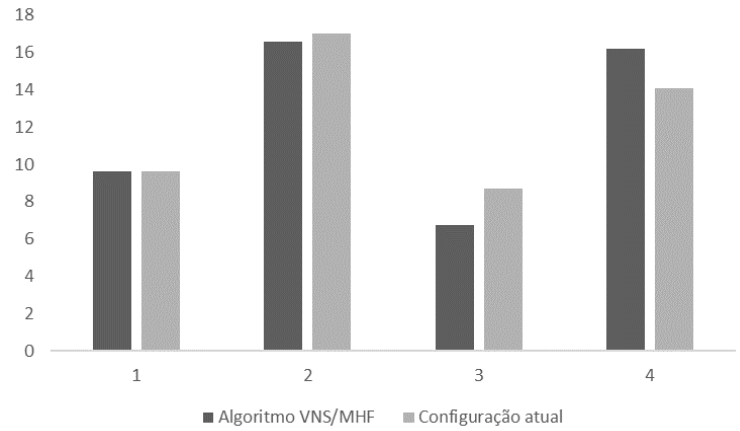

Figura 14: Taxas de ocupação dos servidores utilizando o Algoritmo VNS/MHF

Além de avaliar as taxas de ocupação dos servidores, o tempo médio de viagem no sistema e o tempo de viagem dos servidores também foram analisados. A Tabela 19 apresenta os resultados obtidos utilizando o Algoritmo VNS/MHF e os resultados da configuração atual do SAMU-OP/MA. Foi observada uma redução no tempo médio de viagem no sistema da configuração otimizada em relação à configuração atual. Adicionalmente, foi identificada uma redução no tempo de viagem do servidor 3 e aumento no tempo de viagem dos servidores 2 e 4, parcialmente explicados pela alteração da política de despacho.

Tabela 19 – Tempos médio de viagem utilizando o Algoritmo VNS/MHF

	Tempo médio de viagem no sistema	**Tempo médio de viagem do servidor (TU_i)**			
		1	**2**	**3**	**4**
Configuração atual	13,61	14,46	12,89	16,42	12,86
Algoritmo VNS/MHF	13,42	14,46	13,29	14,70	12,94
Desvio	-1,38%	0,00%	3,06%	-10,50%	0,56%

6.2.2 Análise do SAMU-OP/MA considerando aumento na demanda

A partir dos dados coletados e analisados, foi observado que os sistemas médicos emergenciais estão sujeitos à variação das taxas de chegada dos chamados. Além disso,

situações de aumento de demanda podem ocorrer em função de outros fatores. Por exemplo, foi identificado que devido a Covid-19, em algumas localidades, a demanda por ambulâncias do SAMU aumentou em mais de 130% (REDE GLOBO, 2021).

Sendo assim, foi proposto um cenário alternativo com o intuito de avaliar aumentos de demanda de 20%, 30% e 40% na taxa de chegada de chamados, proporcional a taxa de chegada por localidade e tipo de chamado.

Após a execução do Algoritmo VNS/MHF, foram obtidas as soluções finais apresentadas na Tabela 20. A solução final nos três cenários de aumento de demanda foi igual a solução da configuração otimizada, apresentado na Seção 6.2.1.

Tabela 20 – Soluções encontradas para os cenários de aumento na demanda

Aumento na demanda	Solução final
20% **30%** **40%**	$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$

Adicionalmente, foi avaliado o impacto do aumento da demanda no tempo médio de viagem do sistema e nas taxas de ocupação dos servidores. Como pode ser observado na Figura 15, o aumento na taxa de chegada dos chamados impacta diretamente no tempo médio de viagem do sistema, sendo que, um aumento de 40% da demanda resultou em um aumento de 3,16% no tempo médio de viagem no sistema.

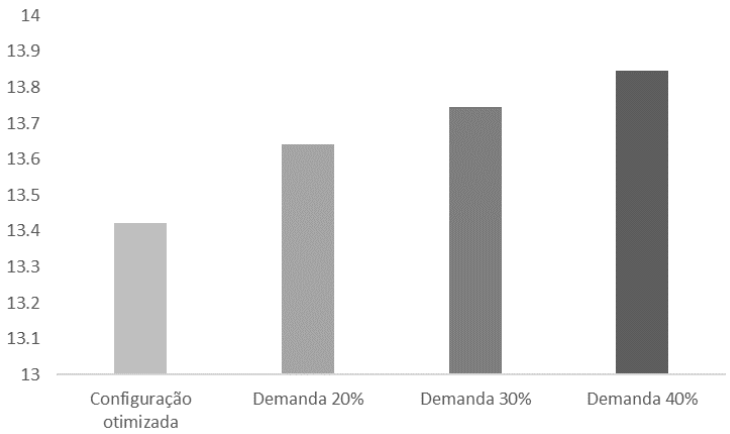

Figura 15: Tempo médio de viagem para os cenários de aumento na demanda

Como esperado, o aumento nas taxas de chegada dos chamados, aumentou a taxa de ocupação dos servidores, conforme ilustrado na Figura 16. Nesse caso, o aumento na taxa de ocupação dos servidores foi, em geral, praticamente proporcional ao aumento na taxa de chegada de chamados. Em casos específicos, como por exemplo, no caso do

servidor 3, foi observado um aumento de 61,34% na taxa ocupação no cenário de aumento de 40% na demanda, em função do servidor 2, preferencial para os chamados provenientes dos átomos 1 e 3 se encontrar mais ocupado, dessa forma, o servidor 3 passou a se deslocar mais frequentemente a esses átomos para realizar o atendimento.

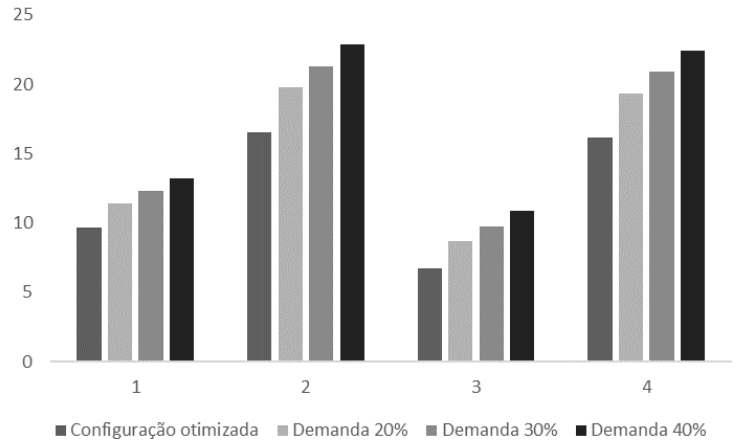

Figura 16: Taxa de ocupação dos servidores nos cenários de aumento na demanda

6.2.3 Análise do SAMU-OP/MA utilizando três servidores

Nos serviços médicos emergenciais, frequentemente, as ambulâncias precisam passar por manutenções, sejam elas de natureza preventiva ou corretiva. Com isso, esse cenário foi proposto com o intuito de avaliar qual a localização dos servidores deve ser utilizada para o caso de uma USB se encontrar indisponível devido a manutenção.

Nesse cenário, as USBs disponíveis foram consideradas homogêneas, ou seja, a taxa de atendimento dos dois servidores disponíveis foi considerada igual. Para tanto, foi realizada a média ponderada das taxas de atendimento dos servidores 2, 3 e 4 do SAMU-OP/MA, apresentadas na Tabela 12.

Após a execução do algoritmo de otimização para o cenário com três servidores disponíveis, uma USA e duas USBs, os servidores foram alocados nos átomos apresentados na Figura 17. Como esperado, a USA foi alocada no átomo 1 e os servidores 2 e 3 foram alocados nas cidades de Mariana e Ouro Preto (átomos 4 e 1), respectivamente, uma vez que, esses átomos apresentam as maiores taxas de chegada de chamados.

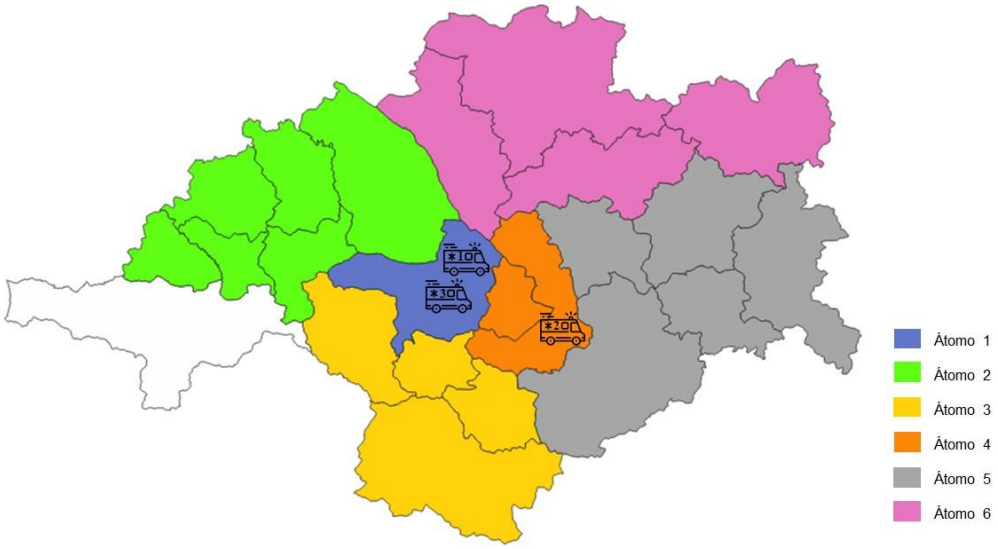

Figura 17: Localização dos três servidores na solução final

Os servidores 2 e 3 absorveram a maioria dos chamados que eram atendidos pela USB que passou a estar indisponível. Dessa forma, foi identificado um aumento na taxa de ocupação desses servidores, conforme apresentado na Figura 18. Vale destacar que em comparação à taxa de ocupação dos servidores básicos, servidores 2, 3 e 4 na configuração otimizada, os servidores 2 e 3 nessa configuração apresentaram, respectivamente, aumentos de 46,70% e 69,05% nas taxas de ocupação. Já o servidor 1 apresentou um aumento de 16,23% na taxa de ocupação em comparação à taxa de ocupação do servidor 1 na configuração otimizada.

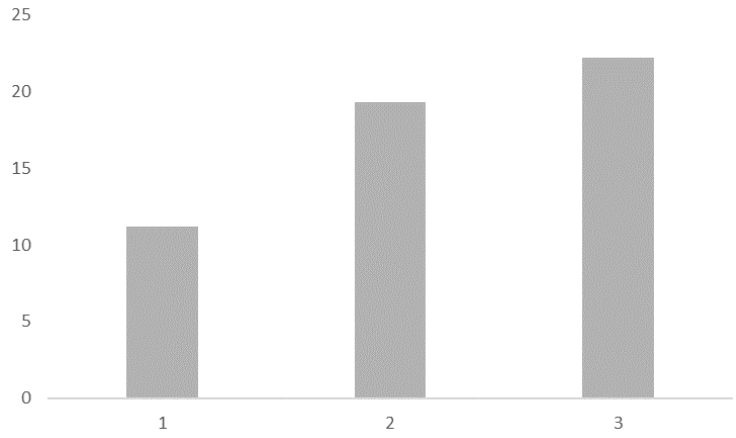

Figura 18: Taxas de ocupação dos três servidores

Na Tabela 21 são apresentados os tempos médio de viagem dos servidores e o tempo médio de viagem no sistema. Traçando um comparativo entre os tempos médio de viagem dos servidores desse cenário e da configuração otimizada do SAMU-OP/MA, foi

identificado que o servidor 1, preferencial para atender os chamados do tipo *a*, praticamente não apresentou alteração em seu tempo médio de viagem. Já os servidores 2 e 3 apresentaram um aumento de 10,82% e 9,89%, em comparação ao tempo médio de viagem dos servidores básicos na configuração otimizada (servidores 2, 3 e 4). É importante ressaltar que foi identificado um aumento de 9,25% no tempo médio de viagem no sistema em comparação com a configuração otimizada.

Tabela 21 – Tempos médio de viagem na análise com três servidores disponíveis

	Tempo médio de viagem no sistema	**Tempo médio de viagem do servidor (TU_i)**		
		1	**2**	**3**
Análise com três servidores	14,66	14,49	14,77	14,64

6.2.4 Análise do SAMU-OP/MA considerando a regulação brasileira

A Portaria nº 1.864/GM, editada pelo Ministério da Saúde em 29 de setembro de 2003, instituiu o componente pré-hospitalar móvel, por intermédio da implantação do SAMU, em todo território brasileiro, no âmbito do SUS (MINISTÉRIO DA SAÚDE, 2003b). Adicionalmente, em seu §3º, a Portaria dispõe que: "As ambulâncias serão adquiridas na proporção de um veículo de suporte básico à vida para cada grupo de 100.000 a 150.000 habitantes, e de um veículo de suporte avançado à vida para cada 400.000 a 450.000 por habitantes".

Segundo o Censo Demográfico de 2010 (IBGE, 2011), Ouro Preto possui uma população total de 70.281 habitantes, enquanto Mariana possui 54.219 habitantes, totalizando uma população de 124.500 habitantes para a região estudada. Seguindo as diretrizes previstas na Portaria nº 1.864/GM, o SAMU-OP/MA deveria contar com uma USB e nenhuma USA. Dessa forma, nesse cenário, foi avaliado como o sistema funcionaria se estive sido configurado obedecendo a portaria que dispõe sobre o dimensionamento de ambulância no Brasil.

Para esse caso, em que apenas um servidor está disponível para a região, o servidor atende tanto os chamados do tipo *a* quanto os chamados do tipo *b*, dessa forma, as taxas de chegada dos chamados do tipo *a* e tipo *b*, mostradas na Tabela 9, foram agregadas na taxa de chegada dos chamados no átomo *i* (λ_i). A taxa de atendimento do servidor (μ) foi obtida a partir da média ponderada das taxas de atendimento dos quatro servidores do SAMU-OP/MA, apresentadas na Tabela 12.

O algoritmo proposto foi utilizado para avaliar o cenário proposto. A Tabela 22 apresenta os tempos médios de viagem no sistema para as seis localizações possíveis do servidor disponível.

Tabela 22 – Tempos médio de viagem para as configurações possíveis de localização do servidor

Solução	Tempo médio de viagem no sistema
[1, 0, 0, 0, 0, 0]	17,96
[0, 1, 0, 0, 0, 0]	30,33
[0, 0, 1, 0, 0, 0]	38,60
[0, 0, 0, 1, 0, 0]	18,64
[0, 0, 0, 0, 1, 0]	33,17
[0, 0, 0, 0, 0, 1]	55,06

Caso o SAMU-OP/MA fosse configurado obedecendo a portaria que dispõe sobre o dimensionamento de ambulância no Brasil, o servidor disponível deveria ser alocado na cidade de Ouro Preto (átomo 1), pois para essa configuração, esse local propicia um menor tempo de viagem no sistema. É importante destacar que nesse cenário, a taxa de ocupação do servidor é de 48,45% e ocorreu um aumento de 33,81% no tempo médio de viagem no sistema em comparação com a configuração atual.

6.2.5 Análise do SAMU-OP/MA considerando regulamentações internacionais

De acordo com Pons e colaboradores (2005), um marcador de qualidade no atendimento de serviços médicos emergenciais é medido pelo cumprimento de uma diretriz de tempo de resposta de 8 minutos.

Com isso, foi proposto esse cenário, como um a tentativa de aproximar o tempo médio de resposta, conforme preconizado internacionalmente. Nesse cenário, assumiu-se que o tempo em fila no sistema é baixo e, por isso, o tempo médio de resposta foi aproximado pelo tempo médio de viagem no sistema.

Para tanto, inicialmente, foram adicionados três servidores do tipo USB ao SAMU-OP/MA. A abordagem proposta foi modificada para avaliar o SAMU-OP/MA utilizando uma USA e seis USBs. Dado que nesse cenário sete servidores devem ser alocados, durante execução do algoritmo, foi definido como critério de parada um tempo máximo igual a 1200 segundos. Após a execução do Algoritmo VNS/MHF para o cenário proposto, os servidores foram alocados seguindo a configuração apresentada na matriz (57).

$$[l_{ik}] = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \quad (57)$$

A configuração apresentada na matriz (56) resultou nos tempos médio de viagem dos servidores e no tempo médio de viagem no sistema apresentado na Tabela 23. Fazendo um comparativo entre os tempos médio de viagem dos servidores desse cenário e da configuração otimizada do SAMU-OP/MA, foi identificado que o servidor 1, preferencial para atender os chamados do tipo *a*, praticamente não apresentou alteração em seu tempo médio de viagem. Já os servidores 2, 3, 4, 5, 6 e 7 apresentaram uma redução de 18,30%, 21,94%, 23,54%, 16,70%, 20,92% e 16,95%, em comparação ao tempo médio de viagem dos servidores básicos na configuração otimizada (servidores 2, 3 e 4).

Tabela 23 – Tempos médio de viagem na análise com sete servidores disponíveis

	Tempo médio de viagem no sistema	**Tempo médio de viagem do servidor (TU_i)**						
		1	**2**	**3**	**4**	**5**	**6**	**7**
Regulamentação internacional	11,36	14,46	10,89	10,40	10,19	11,10	10,54	11,07

Vale destacar que foi identificado uma redução de 15,38% no tempo médio de viagem no sistema em comparação com a configuração otimizada. Entretanto, mesmo com a inclusão de três do tipo USB no sistema, não foi possível atender o tempo médio de resposta igual ou inferior a 8 minutos, conforme preconizado pela legislação internacional.

O não atendimento à legislação internacional se deve ao fato dos municípios de Ouro Preto e Mariana possuírem uma zona rural extensa e pela necessidade de realizar o agrupamento dos distritos desses municípios em átomos geográficos, devido à baixa taxa de chegada de chamados provenientes dos distritos (Figura 7). Com o agrupamento das localidades em átomos, o tempo de viagem dentro de um mesmo átomo geográfico foi definido como 10 minutos, conforme apresentado na Tabela 10. Esse tempo considerou o tamanho dos átomos e a dificuldade de deslocamento dentro dos átomos. Nesse caso, seria necessário utilizar átomos menores para atingir os tempos preconizados internacionalmente. Entretanto, para atingir esses tempos acredita-se que será necessário acrescentar USAs e USBs no sistema.

6.2.6 Análise de cenário fictício de maior porte

Com o intuito de observar como a abordagem desenvolvida se comporta na solução de problemas de maior escala, foi gerada uma instância com 10 servidores e 100 átomos. Foram utilizados uma USA, como servidor exclusivamente dedicado aos chamados do tipo *a*, e nove USBs, como servidores preferenciais para os chamados do tipo *b* e *backup* para chamados do tipo *a*. Na geração da matriz l_{ik} inicial, cada um dos 10 servidores foi alocado aleatoriamente em um dos 100 átomos disponíveis. Dessa forma, o algoritmo de geração de instância construído permitiu a alocação de mais de um servidor em um átomo da região avaliada.

Após gerar a matriz l_{ik} inicial, o algoritmo proposto preencheu a matriz τ_{kj}, com valores uniformemente distribuídos entre 5 e 30, ou seja, o menor tempo de viagem entre os átomos foi igual a 5 minutos e o maior tempo de viagem entre esses átomos foi igual a 30 minutos. Vale destacar que o tempo de deslocamento dentro dos átomos foi definido como sendo 5 minutos.

Em seguida, foram atribuídas taxas de chegadas dos chamados aos 100 átomos do sistema e as taxas de atendimento dos servidores. As taxas de chegadas dos chamados do tipo *a* foram distribuídas uniformemente entre 0,001 e 0,01 e as taxas de chegadas dos chamados do tipo *b* entre 0,01 e 0,1 nos átomos. A taxa de atendimento do servidor preferencial para os chamados do tipo *a* foi definida como 0,7 e os servidores preferenciais para os chamados do tipo *b* foram considerados homogêneos, com taxa de atendimento igual a 0,9.

Definidos os parâmetros de entrada do modelo, a abordagem desenvolvida foi modificada para representar um sistema de 10 servidores e 100 átomos de demanda. Dado que nesse cenário, 10 servidores devem ser alocados, para atender 100 átomos, durante a execução do algoritmo, foi definido como critério de parada um tempo máximo igual a 60 minutos. Após a execução do Algoritmo VNS/MHF para o cenário proposto, foram obtidos os resultados apresentados na Tabela 24. Conforme ilustrado na Tabela 24, o tempo médio de viagem no sistema da configuração otimizada foi 5,09% inferior ao tempo médio de viagem da configuração inicial.

Tabela 24 – Tempos médio de viagem no cenário fictício de maior porte

	Tempo médio de viagem no sistema
Configuração inicial	12,52
Configuração otimizada	11,88

Adicionalmente, foi realizado um comparativo entre os tempos médio de viagem dos servidores na configuração inicial e a configuração otimizada. Como pode ser observado na Figura 19, todos os servidores, com exceção do servidor 1 preferencial dos chamados do tipo *a*, apresentaram uma redução em seu tempo médio de viagem.

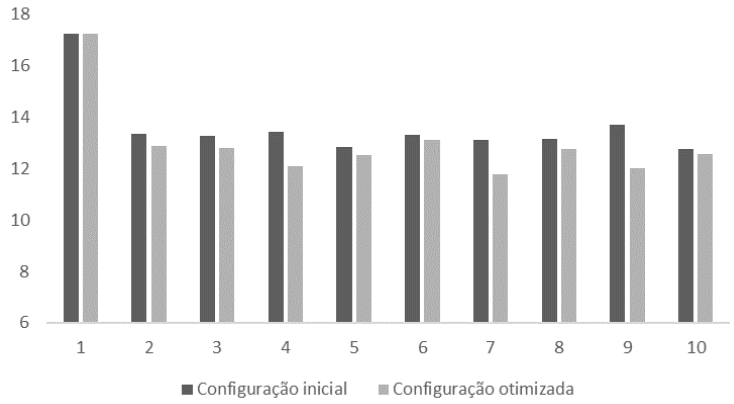

Figura 19: Tempos médio de viagem dos servidores no cenário de maior porte

O Algoritmo VNS/MHF mostrou-se uma ferramenta apropriada e flexível para analisar e otimizar serviços emergenciais de maior escala.

7 Conclusão e Perspectivas Futuras

Essa dissertação aprofunda os estudos sobre o SAMU de Ouro Preto e Mariana. Conforme mencionado, nos serviços médicos emergenciais, a rapidez na resposta ao chamado é um fator crucial para a sobrevida do paciente. Dessa forma, foi apresentada uma abordagem para análise, planejamento e otimização da localização dos servidores do SAMU OP-MA.

Esse trabalho apresentou o Algoritmo VNS/MHF que incorpora o MHF à meta-heurística VNS, permitindo considerar as particularidades das políticas de despacho (*backup* parcial) e localização dos servidores. Adicionalmente, foi demonstrado que essa abordagem pode ser utilizada para analisar as medidas de desempenho SAMU de Ouro Preto e Mariana, assim como, minimizar o tempo médio de viagem no sistema.

Os resultados obtidos a partir da implementação do Algoritmo VNS/MHF impactaram positivamente no desempenho do sistema, especialmente no tempo médio de viagem no sistema. Foi demonstrado que uma política de despachos que considera a menor distância entre os átomos resulta em um tempo médio de viagem no sistema inferior à política atual de despachos de servidores utilizada pelo SAMU-OP/MA.

Conforme mencionado, os resultados da abordagem desenvolvida mostraram um menor tempo de viagem no sistema. Entretanto, comparando com o sistema na configuração atual não foi sugerida mudança de localização de servidores, com exceção da transferência troca de localização dos servidores 2 e 4, o que demonstra a adequação da localização atual das ambulâncias do SAMU-OP/MA, considerando as restrições orçamentárias.

Ademais, foram avaliados cenários utilizando configurações alternativas para o sistema. Foi observado que os sistemas médicos emergenciais estão sujeitos à variação das taxas de chegada dos chamados. Sendo assim, foi proposto um cenário em que a demanda pelos atendimentos aumentou proporcionalmente à taxa de chegada por localidade e tipo de chamado. Os resultados desse cenário corroboram que o SAMU-OP/MA consegue fornecer um serviço com tempo médio de viagem no sistema próximo ao tempo médio de viagem atual, mesmo com um aumento de 40% na taxa de chegada dos chamados.

Dado que nos serviços médicos emergenciais, geralmente, as ambulâncias precisam passar por manutenções, foi proposto um cenário alternativo visando avaliar qual a localização dos servidores deve ser utilizada para o caso de uma USB se encontrar indisponível devido a manutenção. A implementação do Algoritmo VNS/MHF nesse cenário demonstrou que o servidor 1, preferencial para os chamados do tipo *a*, deve permanecer alocado no átomo 1 e os servidores 2 e 3 devem ser alocados nas cidades de Mariana e Ouro Preto (átomos 4 e 1).

Em uma primeira tentativa de analisar o sistema utilizando as regulamentações específicas, brasileira e internacional, foi avaliado qual deveria ser a configuração do SAMU-OP/MA, caso o sistema possuísse apenas uma USB, conforme disposto no §3º da Portaria nº 1.864/GM. Os resultados da abordagem construída demonstram que esse servidor deveria ser localizado na cidade de Ouro Preto e apresentaria um tempo médio de viagem igual a 17,96 minutos. Esse tempo médio de viagem é muito superior ao tempo de 8 minutos preconizado pela legislação internacional, devido ao fato dos municípios de Ouro Preto e Mariana apresentarem uma zona rural com grande extensão territorial. Além disso, os resultados obtidos, como esperado, são superiores aos praticados pelo sistema na configuração atual. Dessa forma, para a realidade dos municípios estudados, os resultados mostraram um descompasso entre a regulamentação brasileira e a internacional.

Visando analisar o sistema considerando as regulamentações internacionais, foram adicionados três servidores do tipo USB ao SAMU-OP/MA. Dessa forma, o Algoritmo VNS/MHF foi modificado para analisar essa nova configuração e apresentou um tempo médio de viagem no sistema igual a 11,36 minutos, resultando em uma redução de 15,38% em comparação à configuração do SAMU-OP/MA otimizada. No entanto, não foi possível atender a regulamentação internacional, devido à necessidade de agrupar as localidades estudadas em átomos geográficos.

Por fim, foi proposto um cenário fictício de maior porte para avaliar o método desenvolvido. A abordagem construída foi implementada no cenário proposto e se mostrou adequada para analisar e otimizar serviços emergenciais de maior porte, uma vez que o tempo médio de viagem no sistema da configuração otimizada do Algoritmo VNS/MHF foi 5,09% inferior ao tempo médio de viagem no sistema da configuração inicial.

Como perspectivas de trabalhos futuros, pretende-se estender o método desenvolvido para tratar os objetivos de minimizar o tempo de resposta e minimizar os despachos interdistritos. Pretende-se também utilizar o método desenvolvido para subsidiar a análise das portarias de dimensionamento de servidores emergenciais no Brasil, contribuindo na formulação de políticas públicas.

REFERÊNCIAS

AHMADI-JAVID, Amir; SEYEDI, Pardis; SYAM, Siddhartha S. A survey of healthcare facility location. **Computers and Operations Research**, *[S. l.]*, v. 79, p. 223–263, 2017. DOI: 10.1016/j.cor.2016.05.018.

AKDOGAN, M. Altan; BAYINDIR, Z. Pelin; IYIGUN, Cem. Location analysis of emergency vehicles using an approximate queueing model. **Transportation Research Procedia**, *[S. l.]*, v. 22, p. 430–439, 2017. DOI: 10.1016/j.trpro.2017.03.018.

AKDOGAN, M. Altan; BAYINDIR, Z. Pelin; IYIGUN, Cem. Locating emergency vehicles with an approximate queuing model and a meta-heuristic solution approach ☆. **Transportation Research Part C**, *[S. l.]*, v. 90, n. March 2017, p. 134–155, 2018. DOI: 10.1016/j.trc.2018.01.014.

ANOKIC, Ana; STANIMIROVIC, Zorica; DAVIDOVIC, Tatjana; STAKIC, Dorde. Variable neighborhood search based approaches to a vehicle scheduling problem in agriculture. **International Transactions in Operational Research**, *[S. l.]*, v. 27, p. 26–56, 2020. DOI: 10.1111/itor.12480.

ANSARI, Sardar; MCLAY, Laura Albert; MAYORGA, Maria E. A Maximum Expected Covering Problem for District Design. **Transportation Science**, *[S. l.]*, n. Articles in Advance, p. 1–55, 2015. DOI: 10.1287/trsc.2015.0610.

ANSARI, Sardar; YOON, Soovin; ALBERT, Laura A. An approximate hypercube model for public service systems with co-located servers and multiple response. **Transportation Research Part E**, *[S. l.]*, v. 103, p. 143–157, 2017. DOI: 10.1016/j.tre.2017.04.013.

ARMAS, Jesica De; MELIÁN-BATISTA, Belén. Variable Neighborhood Search for a Dynamic Rich Vehicle Routing Problem with time windows q. **COMPUTERS & INDUSTRIAL ENGINEERING**, *[S. l.]*, v. 85, p. 120–131, 2015. DOI: 10.1016/j.cie.2015.03.006.

ATKINSON, J. B.; KOVALENKO, I. N.; KUZNETSOV, N.; MYKHALEVYCH, K. V. A hypercube queueing loss model with customer-dependent service rates. **European Journal of Operational Research**, *[S. l.]*, v. 191, n. 1, p. 221–237, 2008. DOI: 10.1016/j.ejor.2007.08.014.

BARNEVELD, T. C. Van; MEI, R. D. Van Der; BHULAI, S. Compliance tables for an EMS system with two types of medical response units. **Computers and Operations Research**, *[S. l.]*, v. 80, p. 68–81, 2017. DOI: 10.1016/j.cor.2016.11.013.

BATTA, Rajan; DOLAN, June M.; KRISHNAMURTHY, Nirup N. The Maximal

Expected Covering Problem : Revisited. **Transportation Science**, *[S. l.]*, v. 23, n. 4, p. 277–287, 1989.

BÉLANGER, V.; RUIZ, A.; SORIANO, P. Recent optimization models and trends in location , relocation , and dispatching of emergency medical vehicles. **European Journal of Operational Research**, *[S. l.]*, v. 272, p. 1–23, 2019. DOI: 10.1016/j.ejor.2018.02.055.

BERLIN, N.; LIEBMAN, J. O. N. C. MATHEMATICAL ANALYSIS OF EMERGENCY AMBULANCE LOCATION *. **Socio-Economic Planning Sciences**, *[S. l.]*, n. 1, p. 323–328, 1974.

BLANCHARD, Ian E. et al. Emergency Medical Services Response Time and Mortality in an Urban Setting. **Prehospital Emergency Care**, *[S. l.]*, v. 3127, n. March 2016, 2012. DOI: 10.3109/10903127.2011.614046.

BLANK, Felix. A spatial queuing model for the location decision of emergency medical vehicles for pandemic outbreaks: the case of Za'atari refugee camp. **Journal of Humanitarian Logistics and Supply Chain Management**, *[S. l.]*, 2021. DOI: 10.1108/JHLSCM-07-2020-0058.

BOYACI, Burak; GEROLIMINIS, Nikolas. Approximation methods for large-scale spatial queueing systems. **Transportation Research Part B: Methodological**, *[S. l.]*, v. 74, p. 151–181, 2015. DOI: 10.1016/j.trb.2014.12.011.

BRIMBERG, Jack; MLADENOVIĆ, Nenad; TODOSIJEVIĆ, Raca; UROŠEVIĆ, Dragan. General variable neighborhood search for the uncapacitated single allocation p - hub center problem. **Optimization Letters**, *[S. l.]*, v. 11, n. 2, p. 377–388, 2016. DOI: 10.1007/s11590-016-1004-x.

BROTCORNE, Luce; LAPORTE, Gilbert; SEMET, Frédéric. Ambulance location and relocation models. **European Journal of Operational Research**, *[S. l.]*, v. 147, p. 451–463, 2003. DOI: 10.1016/S0377-2217(02)00364-8.

BUDGE, Susan; INGOLFSSON, Armann; ERKUT, Erhan. Approximating Vehicle Dispatch Probabilities for Emergency Service Systems with Location-Specific Service Times and Multiple Units per Location Approximating Vehicle Dispatch Probabilities for Emergency Service Systems with Location-Specific Service Times. **Operations Research**, *[S. l.]*, v. 57, p. 251–255, 2009. DOI: 10.1287/opre.1080.0591.

BURWELL, Timothy H.; JARVIS, James P.; MCKNEW, Mark A. Modeling co-located servers and dispatch ties in the hypercube model. **Computers and Operations Research**, *[S. l.]*, v. 20, n. 2, p. 113–119, 1993. DOI: 10.1016/0305-0548(93)90067-S.

BURWELL, Timothy H.; MCKNEW, Mark A.; JARVIS, James P. An application of a spatially distributed queuing model to an ambulance system. **Socio-Economic Planning Sciences**, *[S. l.]*, v. 26, n. 4, p. 289–300, 1992. DOI: 10.1016/0038-0121(92)90005-P.

CHAMPMAN, S. C.; WHITE, J. A. Probabilistic formulations of emergency service facilities location problems. *In*: ORSA/TIMS CONFERENCE, SAN JUAN, PUERTO RICO 1974, **Anais** [...]. [s.l: s.n.]

CHELST, Kenneth R.; BARLACH, Ziv. Multiple Unit Dispatches in Emergency Services : Models to Estimate System Performance. *[S. l.]*, n. September 2015, 1981.

CHIYOSHI, Fernando; GALVÃO, Roberto D.; MORABITO, Reinaldo. O uso do modelo hipercubo na solução de problemas de localização probabilísticos. **Gestão & Produção**, *[S. l.]*, v. 7, n. 2, p. 146–174, 2000. DOI: 10.1590/s0104-530x2000000200005.

CHIYOSHI, Fernando Y.; GALVÃO, Roberto D.; MORABITO, Reinaldo. A note on solutions to the maximal expected covering location problem. **Computers and Operations Research**, *[S. l.]*, v. 30, n. 1, p. 87–96, 2003. DOI: 10.1016/S0305-0548(01)00083-1.

CHURCH, Richard; REVELLE, Charles. THE MAXIMAL COVERING LOCATION PROBLEM. **Regional Science Association**, *[S. l.]*, 1974.

COSTA, Leandro R.; ALOISE, Daniel; MLADENOVI, Nenad. Less is more : basic variable neighborhood search heuristic for balanced minimum sum-of-squares clustering. **Information Sciences**, *[S. l.]*, v. 416, p. 247–253, 2017. DOI: 10.1016/j.ins.2017.06.019.

DASKIN, Mark S. Application of an Expected Covering Model to Emergency Medical Service System Design. **Decision Science**, *[S. l.]*, v. 13, n. 3, p. 416–439, 1982. DOI: 10.1111/j.1540-5915.1982.tb00159.x.

DASKIN, Mark S.; STERN, Edmund H. A Hierarchical Objective Set Covering Model for Emergency Medical Service Vehicle Deployment. **Transportation Science**, *[S. l.]*, v. 15, n. 2, p. 85–166, 1981. DOI: https://doi.org/10.1287/trsc.15.2.137.

DE SOUZA, Regiane Máximo; MORABITO, Reinaldo; CHIYOSHI, Fernando Y.; IANNONI, Ana Paula. 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**, *[S. l.]*, v. 242, n. 1, p. 274–285, 2015. DOI: 10.1016/j.ejor.2014.09.056.

FATTAHI, Parviz; KENESHLOO, Sanaz; DANESHAMOOZ, Fatemeh; AHMADI, Samad. A Hybrid Genetic Algorithm and Parallel Variable Neighborhood Search for Job

Shop Scheduling with an Assembly Stage. **International Journal of Industrial Engineering & Production Research**, *[S. l.]*, v. 30, n. 1, p. 25–37, 2019. DOI: 10.22068/ijiepr.30.1.25.

GALVÃO, Roberto D.; MORABITO, Reinaldo. Emergency service systems: The use of the hypercube queueing model in the solution of probabilistic location problems. **International Transactions in Operational Research**, *[S. l.]*, v. 15, n. 5, p. 525–549, 2008. DOI: 10.1111/j.1475-3995.2008.00654.x.

GEROLIMINIS, Nikolas; KARLAFTIS, Matthew G.; SKABARDONIS, Alexander. A spatial queuing model for the emergency vehicle districting and location problem. **Transportation Research Part B: Methodological**, *[S. l.]*, v. 43, n. 7, p. 798–811, 2009. DOI: 10.1016/j.trb.2009.01.006.

GEROLIMINIS, Nikolas; KEPAPTSOGLOU, Konstantinos; KARLAFTIS, Matthew G. A hybrid hypercube - Genetic algorithm approach for deploying many emergency response mobile units in an urban network. **European Journal of Operational Research**, *[S. l.]*, v. 210, n. 2, p. 287–300, 2011. DOI: 10.1016/j.ejor.2010.08.031.

GOLABIAN, Hamed; ARKAT, Jamal; FARUGHI, Hiwa; TAVAKKOLI-MOGHADDAM, Reza. A simulation-optimization algorithm for return strategies in emergency medical systems. **Simulation: Transactions of the Society for Modeling and Simulation International 1–24**, *[S. l.]*, p. 1–24, 2021. DOI: 10.1177/00375497211006175.

GUAN, Jian; LIN, Geng. Hybridizing variable neighborhood search with ant colony optimization for solving the single row facility layout problem. **European Journal of Operational Research**, *[S. l.]*, 2015. DOI: 10.1016/j.ejor.2015.08.014.

HAKIMI, S. L. Optimum Distribution of Switching Centers in a Communication Network and Some Related Graph Theoretic Problems. **Operations Research**, *[S. l.]*, n. August 2015, 1965.

HANSEN, Pierre; MLADENOVIĆ, Nenad; BRIMBERG, Jack; PÉREZ, José A. Moreno. Variable Neighborhood Search. *In*: **Handbook of Metaheuristics**. Third Edit ed. [s.l.] : Springer, 2019. p. 57–97.

IANNONI, Ana Paula. **Otimização da configuração e operação de sistemas médico emergencial em rodovias utilizando o modelo hipercubo**. 2005. Universidade Federal de São Carlos, *[S. l.]*, 2005.

IANNONI, Ana Paula; MORABITO, Reinaldo. A multiple dispatch and partial backup hypercube queuing model to analyze emergency medical systems on highways.

Transportation Research Part E, *[S. l.]*, v. 43, p. 755–771, 2007. DOI: 10.1016/j.tre.2006.05.005.

IANNONI, Ana Paula; MORABITO, Reinaldo. Otimização da localização das bases de ambulâncias e do dimensionamento das suas regiões de cobertura em rodovias. **Produção**, *[S. l.]*, v. 18, n. 1, p. 47–63, 2008. DOI: 10.1590/S0103-65132008000100004.

IANNONI, Ana Paula; MORABITO, Reinaldo; SAYDAM, Cem. A hypercube queueing model embedded into a genetic algorithm for ambulance deployment on highways. **Annals of Operations Research**, *[S. l.]*, p. 207–224, 2008. DOI: 10.1007/s10479-007-0195-z.

IANNONI, Ana Paula; MORABITO, Reinaldo; SAYDAM, Cem. An optimization approach for ambulance location and the districting of the response segments on highways. **European Journal of Operational Research**, *[S. l.]*, v. 295, p. 528–542, 2009. DOI: 10.1016/j.ejor.2008.02.003.

IANNONI, Ana Paula; MORABITO, Reinaldo; SAYDAM, Cem. Optimizing large-scale emergency medical system operations on highways using the hypercube queuing model. **Socio-Economic Planning Sciences**, *[S. l.]*, v. 45, n. 3, p. 105–117, 2011. DOI: 10.1016/j.seps.2010.11.001.

IBGE. **Censo Demográfico de 2010**. 2011. Disponível em: https://censo2010.ibge.gov.br/sinopseporsetores/?nivel=st. Acesso em: 29 mar. 2020.

KALAYCI, Can B.; KAYA, Can. An ant colony system empowered variable neighborhood search algorithm for the vehicle routing problem with simultaneous pickup and delivery. **Expert Systems With Applications**, *[S. l.]*, v. 66, p. 163–175, 2016. DOI: 10.1016/j.eswa.2016.09.017.

KIM, Sun Hoon; LEE, Young Hoon. Iterative optimization algorithm with parameter estimation for the ambulance location problem. **Health Care Management Science**, *[S. l.]*, v. 19, n. 4, p. 362–382, 2016. DOI: 10.1007/s10729-015-9332-4.

LARSON, Richard C. A hypercube queuing model for facility location and redistricting in urban emergency services. **Computers and Operations Research**, *[S. l.]*, v. 1, n. 1, p. 67–95, 1974. DOI: 10.1016/0305-0548(74)90076-8.

LARSON, Richard C.; MCKNEW, Mark A. Police Patrol-Initiated Activities Within a Systems Queueing Model. **Management Science**, *[S. l.]*, v. 28, n. 7, p. 759–774, 1982.

LARSON, Richard C.; ODONI, Amedeo R. **Urban operations research**. [s.l.] : Dynamic Ideas; 2nd Edition, 2007.

LEI, Deming. Computers & Industrial Engineering Variable neighborhood search for

two-agent flow shop scheduling. **COMPUTERS & INDUSTRIAL ENGINEERING**, *[S. l.]*, v. 80, p. 125–131, 2015. DOI: 10.1016/j.cie.2014.11.024.

LI, Xueping; ZHAO, Zhaoxia; ZHU, Xiaoyan; WYATT, Tami. Covering models and optimization techniques for emergency response facility location and planning: A review. **Mathematical Methods of Operations Research**, *[S. l.]*, v. 74, n. 3, p. 281–310, 2011. DOI: 10.1007/s00186-011-0363-4.

MARINAKIS, Yannis; MIGDALAS, Athanasios; SIFALERAS, Angelo. A hybrid Particle Swarm Optimization – Variable Neighborhood Search algorithm for Constrained Shortest Path problems. **European Journal of Operational Research**, *[S. l.]*, v. 261, n. 3, p. 819–834, 2017. DOI: 10.1016/j.ejor.2017.03.031. Disponível em: http://dx.doi.org/10.1016/j.ejor.2017.03.031.

MARTINS, Pedro Paulo Scremin; PRADO, Marta Lenise Do. Enfermagem e serviço de atendimento pré-hospitalar: descaminhos e perspectivas. **Revista Brasileira de Enfermagem**, *[S. l.]*, v. 56, n. 1, p. 71–75, 2003. DOI: 10.1590/S0034-71672003000100015.

MAYORGA, Maria E.; BANDARA, Damitha; MCLAY, Laura A. Districting and dispatching policies for emergency medical service systems to improve patient survival. **IIE Transactions on Healthcare Systems Engineering**, *[S. l.]*, p. 39–56, 2013. DOI: 10.1080/19488300.2012.762437.

MCLAY, Laura A. A maximum expected covering location model with two types of servers A maximum expected covering location model with two types of servers. **IIE Transactions**, *[S. l.]*, n. May 2013, p. 37–41, 2009. DOI: 10.1080/07408170802702138.

MENDES, Nilson Felipe Matos; SANTOS, André Gustavo. A Hypercube Queuing Model Approach to the Police Units Allocation Problem A Hypercube Queuing Model Approach to the Police Units Allocation Problem. **In Proceedings of the 18th International Conference on Enterprise Information Systems (ICEIS 2016)**, *[S. l.]*, v. 2, p. 7–81, 2016. DOI: 10.5220/0005837800700081.

MENDONÇA, Fc C.; MORABITO, R. Analysing emergency medical service ambulance deployment on a brazilian highway using thehypercube model. **Journal of the Operational Research Society**, *[S. l.]*, v. 52, n. 3, p. 261–270, 2001. DOI: 10.1057/palgrave.jors.2601097.

MENÉNDEZ, Borja; PARDO, Eduardo G.; ALONSO-AYUSO, Antonio; MOLINA, Elisenda; DUARTE, Abraham. Computers & Operations Research Variable Neighborhood Search strategies for the Order Batching Problem. **Computers and**

Operation Research, *[S. l.]*, p. 1–13, 2016. DOI: 10.1016/j.cor.2016.01.020.

MIGUEL, Paulo A. Cauchick et al. **Metodologia de Pesquisa em Engenharia de Produção e Gestão de Operações**. 3 Edição ed. [s.l.] : Elsevier, 2018.

MINAYO, Maria Cecília De Souza; DESLANDES, Suely Ferreira. Análise da implantação do sistema de atendimento pré-hospitalar móvel em cinco capitais brasileiras. **Cadernos de Saúde Pública**, *[S. l.]*, v. 24, n. 8, p. 1877–1886, 2008. DOI: 10.1590/S0102-311X2008000800016.

MINISTÉRIO DA SAÚDE. **PORTARIA Nº 1863, DE 29 DE SETEMBRO DE 2003**. 2003a. Disponível em: http://bvsms.saude.gov.br/bvs/saudelegis/gm/2003/prt1863_26_09_2003.html. Acesso em: 29 mar. 2020.

MINISTÉRIO DA SAÚDE. **PORTARIA Nº 1.864, DE 29 DE SETEMBRO DE 2003**. 2003b. Disponível em: http://bvsms.saude.gov.br/bvs/saudelegis/gm/2003/prt1864_29_09_2003.html. Acesso em: 29 mar. 2020.

MINISTÉRIO DA SAÚDE. **PORTARIA Nº 1.010, DE 21 DE MAIO DE 2012**. 2012. Disponível em: http://bvsms.saude.gov.br/bvs/saudelegis/gm/2012/prt1010_21_05_2012.html. Acesso em: 29 mar. 2020.

MINISTÉRIO DA SAÚDE. **Serviço de Atendimento Móvel de Urgência (SAMU 192)**. 2019. Disponível em: https://www.saude.gov.br/saude-de-a-z/servico-de-atendimento-movel-de-urgencia-samu-192. Acesso em: 29 mar. 2020.

MLADENOVIĆ, N.; HANSEN, Pierre. VARIABLE NEIGHBORHOOD SEARCH. **Computers and Operations Research**, *[S. l.]*, v. 24, n. 1, p. 1097–1100, 1997.

MLADENOVIĆ, Nenad; TODOSIJEVIĆ, Raca; UROŠEVIĆ, Dragan. Less is more : Basic variable neighborhood search for minimum differential dispersion problem. **Information Sciences**, *[S. l.]*, v. 326, p. 160–171, 2016. DOI: 10.1016/j.ins.2015.07.044.

MOBIN, M.; LI, Z.; AMIRI, M. Performance Evaluation of Tehran-Qom Highway Emergency Medical Service System Using Hypercube Queuing Model. *In*: IIE ANNUAL CONFERENCE. PROCEEDINGS 2015, **Anais** [...]. [s.l: s.n.] p. 1175.

MOON, Sungbae; RYOO, Hyun Wook; AHN, Jae Yun; LEE, Dong Eun; SHIN, Sang Do; PARK, Jeong Ho. ur l P re of. **American Journal of Emergency Medicine**, *[S. l.]*, 2020. DOI: 10.1016/j.ajem.2020.05.108.

MORABITO, Reinaldo; CHIYOSHI, Fernando; GALVÃO, Roberto D. Non-

homogeneous servers in emergency medical systems : Practical applications using the hypercube queueing model. **Socio-Economic Planning Sciences**, *[S. l.]*, v. 42, p. 255–270, 2008. DOI: 10.1016/j.seps.2007.04.002.

PONS, Peter T.; HAUKOOS, Jason S.; BLUDWORTH, Whitney; CRIBLEY, Thomas; PONS, Kathryn A.; MARKOVCHICK, Vincent J. Paramedic response time: Does it affect patient survival? **Academic Emergency Medicine**, *[S. l.]*, v. 12, n. 7, p. 594–600, 2005. DOI: 10.1197/j.aem.2005.02.013.

RAJAGOPALAN, Hari K.; SAYDAM, Cem. A minimum expected response model: Formulation, heuristic solution, and application. **Socio-Economic Planning Sciences**, *[S. l.]*, v. 43, n. 4, p. 253–262, 2009. DOI: 10.1016/j.seps.2008.12.003.

RAJAGOPALAN, Hari K.; SAYDAM, Cem; XIAO, Jing. A multiperiod set covering location model for dynamic redeployment of ambulances. **Computers and Operations Research**, *[S. l.]*, v. 35, n. 3, p. 814–826, 2008. DOI: 10.1016/j.cor.2006.04.003.

RAJAGOPALAN, Hari K.; VERGARA, F. Elizabeth; SAYDAM, Cem; XIAO, Jing. Developing effective meta-heuristics for a probabilistic location model via experimental design. **European Journal of Operational Research**, *[S. l.]*, v. 177, n. 1, p. 83–101, 2007. DOI: 10.1016/j.ejor.2005.11.007.

REDE GLOBO. **Chamados por ambulâncias do Samu aumentam mais de 130% em BH por causa da Covid-19**. 2021. Disponível em: https://g1.globo.com/mg/minas-gerais/noticia/2021/01/12/chamados-por-ambulancias-do-samu-aumentam-mais-de-130percent-em-bh-por-causa-da-covid-19.ghtml.

REVELLE, Charles; HOGAN, Kathleen. The Maximum Availability Location Problem. **Transportation Science**, *[S. l.]*, v. 23, n. 3, p. 192–200, 1989. DOI: 10.1287/trsc.23.3.192.

RODRIGUES, Lásara Fabrícia. **Análise dos Serviços Emergenciais de Manutenção Agrícola e Borracharia na Agroindústria Canavieira Utilizando Teoria de Filas**. 2014. Universidade Federal de São Carlos, São Carlos - SP, *[S. l.]*, 2014.

RODRIGUES, Lásara Fabrícia; MORABITO, Reinaldo; CHIYOSHI, Fernando Y.; IANNONI, Ana Paula; SAYDAM, Cem. Towards hypercube queuing models for dispatch policies with priority in queue and partial backup. **Computers and Operations Research**, *[S. l.]*, v. 84, p. 92–105, 2017. DOI: 10.1016/j.cor.2017.02.021.

RODRIGUES, Lásara Fabrícia; MORABITO, Reinaldo; CHIYOSHI, Fernando Y.; IANNONI, Ana Paula; SAYDAM, Cem. Analyzing an emergency maintenance system in the agriculture stage of a Brazilian sugarcane mill using an approximate hypercube

method. **Computers and Electronics in Agriculture**, *[S. l.]*, v. 151, n. June, p. 441–452, 2018. DOI: 10.1016/j.compag.2018.06.023.

RODRIGUES, Lasara Fabrícia; MORABITO, Reinaldo; SAYDAM, Cem. Queuing analysis of emergency repair systems: A case study of tire repairing in Brazil sugarcane industry. **International Journal of Operations and Quantitative Management**, *[S. l.]*, v. 22, n. 4, p. 335–356, 2016.

RODRIGUEZ, Sebastian A.; DE LA FUENTE, Rodrigo A.; AGUAYO, Maichel M. A facility location and equipment emplacement technique model with expected coverage for the location of fire stations in the Concepción province, Chile. **Computers and Industrial Engineering**, *[S. l.]*, v. 147, n. March, p. 106522, 2020. DOI: 10.1016/j.cie.2020.106522.

SARASOLA, Briseida; DOERNER, Karl F.; SCHMID, Verena; ALBA, Enrique. Variable neighborhood search for the stochastic and dynamic vehicle routing problem. **Annals of Operations Research**, *[S. l.]*, 2015. DOI: 10.1007/s10479-015-1949-7.

SAYDAM, Cem; AYTUĞ, Haldun. Accurate estimation of expected coverage: Revisited. **Socio-Economic Planning Sciences**, *[S. l.]*, v. 37, n. 1, p. 69–80, 2003. DOI: 10.1016/S0038-0121(02)00004-6.

SAYDAM, Cem; REPEDE, John; BURWELL, Timothy. Accurate estimation of expected coverage: A comparative study. **Socio-Economic Planning Sciences**, *[S. l.]*, v. 28, n. 2, p. 113–120, 1994. DOI: 10.1016/0038-0121(94)90010-8.

SICILA, Juan Antonio; QUEMADA, Carlos; ROYO, Beatriz; ESCUÍN, David. An optimization algorithm for solving the rich vehicle routing problem based on Variable Neighborhood Search and Tabu Search metaheuristics. **Journal of Computational and Applied Mathematics**, *[S. l.]*, 2015. DOI: 10.1016/j.cam.2015.03.050.

SIFALERAS, Angelo; KONSTANTARAS, Ioannis; MLADENOVIĆ, Nenad. Variable neighborhood search for the economic lot sizing problem with product returns and recovery. **International Journal of Production Economics**, *[S. l.]*, v. 160, p. 133–143, 2015. DOI: 10.1016/j.ijpe.2014.10.003.

SUDTACHAT, Kanchala; MAYORGA, Maria E.; CHANTA, Sunarin; ALBERT, Laura A. Joint relocation and districting using a nested compliance model for EMS systems. **Computers & Industrial Engineering**, *[S. l.]*, v. 142, p. 106327, 2020. DOI: 10.1016/j.cie.2020.106327.

SWERSEY, Arthur; GOLDRING, Louis; GEYER, Earl D. Improving Fire Department Productivity: Merging Fire and Emergency Medical Units in New Haven. **The Institute**

of Emergency Sciences, *[S. l.]*, 1994.

TAKEDA, Renata Algisi; WIDMER, João A.; MORABITO, Reinaldo. Analysis of ambulance decentralization in an urban emergency medical service using the hypercube queueing model. **Computers and Operations Research**, *[S. l.]*, v. 34, n. 3, p. 727–741, 2007. DOI: 10.1016/j.cor.2005.03.022.

TAKEDA, Renata Algisi; WIDMER, João Alexandre; MORABITO, Reinaldo. Uma Proposta Alternativa Para Avaliação Do Desempenho De Sistemas De Trasnporte Emergencial De Saúde. **Transportes**, *[S. l.]*, v. 9, n. 2, p. 9–26, 2001. DOI: 10.4237/transportes.v9i2.168.

TAKEDA, Renata Algisi; WIDMER, João Alexandre; MORABITO, Reinaldo. Aplicação do modelo hipercubo de filas para avaliar a descentralização de ambulâncias em um sistema urbano de atendimento médico de urgência. **Pesquisa Operacional**, *[S. l.]*, v. 24, n. 1, p. 39–71, 2004. DOI: 10.1590/S0101-74382004000100004.

TODOSIJEVIC, Raca; BENMANSOUR, Rachid; HANAFI, Said; MLADENOVIC, Nenad; ARTIBA, Abdelhakim. Nested general variable neighborhood search for the periodic maintenance problem. **European Journal of Operational Research**, *[S. l.]*, 2016. DOI: 10.1016/j.ejor.2016.01.014.

TOREGAS, Constantine; SWAIN, Ralph; REVELLE, Charles; BERGMAN, Lawrence. The Location of Emergency Service Facilities. **Operations Research**, *[S. l.]*, n. September 2014, 1971. DOI: http://dx.doi.org/10.1287/opre.19.6.1363.

TORO-DÍAZ, Hector; MAYORGA, Maria E.; CHANTA, Sunarin; MCLAY, Laura A. Joint location and dispatching decisions for Emergency Medical Services. **Computers and Industrial Engineering**, *[S. l.]*, v. 64, n. 4, p. 917–928, 2013. DOI: 10.1016/j.cie.2013.01.002.

TORO-DÍAZ, Hector; MAYORGA, Maria E.; MCLAY, Laura A.; RAJAGOPALAN, Hari K.; SAYDAM, Cem. Reducing disparities in large-scale emergency medical service systems. **Journal of the Operational Research Society**, *[S. l.]*, v. 66, n. 7, p. 1169–1181, 2015. DOI: 10.1057/jors.2014.83.

TRUJILLO, Leonardo; ÁLVAREZ-HERNÁNDEZ, Guadalupe; MALDONADO, Yazmin; VERA, Carlos. Comparative analysis of relocation strategies for ambulances in the city of. **Computers in Biology and Medicine**, *[S. l.]*, v. 116, n. September 2019, p. 103567, 2020. DOI: 10.1016/j.compbiomed.2019.103567.

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, *[S. l.]*, v. 80, p. 68–81, 2017. DOI: 10.1016/j.cor.2016.11.013.

WANG, Xingyin; GOLDEN, Bruce; WASIL, Edward. A Steiner Zone Variable Neighborhood Search Heuristic for the Close-Enough Traveling Salesman Problem. **Computers and Operations Research**, *[S. l.]*, v. 101, p. 200–219, 2019. DOI: 10.1016/j.cor.2018.07.023.

YOON, Soovin; ALBERT, Laura A. A dynamic ambulance routing model with multiple response. **Transportation Research Part E**, *[S. l.]*, v. 133, p. 101807, 2020. DOI: 10.1016/j.tre.2019.11.001.

ZAFFAR, Muhammad Adeel; RAJAGOPALAN, Hari K.; SAYDAM, Cem; MAYORGA, Maria; SHARER, Elizabeth. Coverage , survivability or response time: A comparative study of performance statistics used in ambulance location models via simulation–optimization. **Operations Research for Health Care**, *[S. l.]*, v. 11, p. 1–12, 2016. DOI: 10.1016/j.orhc.2016.08.001. Disponível em: http://dx.doi.org/10.1016/j.orhc.2016.08.001.