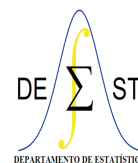




UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE ESTATÍSTICA
BACHARELADO EM ESTATÍSTICA



Jucimara Nunes Quirino

**Otimização em Redes de Filas através de
processamento duplo via NSGA-II + Estratégia gulosa**

Ouro Preto-MG
2026

JUCIMARA NUNES QUIRINO

OTIMIZAÇÃO EM REDES DE FILAS ATRAVÉS DE
PROCESSAMENTO DUPLO VIA NSGA-II + ESTRATÉGIA GULOSA

Monografia de Graduação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas e Biológicas da Universidade Federal de Ouro Preto como requisito parcial para a obtenção do grau de bacharel em Estatística.

Orientador: Anderson Ribeiro Duarte

OURO PRETO

2026

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

Q8o Quirino, Jucimara Nunes.
Otimização em Redes de Filas através de processamento duplo via
NSGA-II + Estratégia gulosa. [manuscrito] / Jucimara Nunes Quirino. -
2026.

24 f.: il.: gráf., tab..

Orientador: Prof. Dr. Anderson Ribeiro Duarte.
Monografia (Bacharelado). Universidade Federal de Ouro Preto.
Instituto de Ciências Exatas e Biológicas. Graduação em Estatística .

1. Rede de filas. 2. Teoria das filas. 3. Filas finitas. 4. Algoritmos
genéticos. I. Duarte, Anderson Ribeiro. II. Universidade Federal de Ouro
Preto. III. Título.

CDU 519.872

Bibliotecário(a) Responsável: Renata Mara de Almeida - CRB-7: 6328



MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
COLEGIADO DO CURSO DE ESTATÍSTICA



FOLHA DE APROVAÇÃO

Jucimara Nunes Quirino

Otimização em Redes de Filas através de processamento duplo via NSGA-II + Estratégia gulosa

Monografia apresentada ao Curso de Estatística da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Estatística

Aprovada em 12 de fevereiro de 2026

Membros da banca

Dr. Anderson Ribeiro Duarte - Orientador (Universidade Federal de Ouro Preto)
Dr. Helgem de Souza Ribeiro Martins (Universidade Federal de Ouro Preto)
Dr. Gabriel Lima de Souza (Universidade Federal de Ouro Preto)

Professor Dr. Anderson Ribeiro Duarte, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 12/02/2026



Documento assinado eletronicamente por **Anderson Ribeiro Duarte, PROFESSOR DE MAGISTERIO SUPERIOR**, em 12/02/2026, às 20:20, 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 **1058644** e o código CRC **BD943627**.

Agradecimentos

Agradeço, primeiramente, a Deus, por sempre me guiar e fortalecer ao longo de toda esta trajetória. Sem Sua presença, nada disso seria possível.

A conclusão deste trabalho representa o encerramento de uma etapa profundamente sonhada e desejada. Trata-se da conquista de um sonho sustentado por muito, esforço e sobretudo, resiliência, valores que aprendi e sempre vi refletidos em meus pais, Geraldo Cirilo Quirino e Geralda das Dores Nunes Quirino, e em meu irmão, Wilian Aparecido Quirino, que nunca deixaram de acreditar em mim, em meus sonhos e nos caminhos que escolhi seguir, oferecendo apoio, incentivo e confiança em todos os momentos desta caminhada.

Agradeço profundamente ao meu professor e orientador, Dr. Anderson Ribeiro Duarte, primeiramente, como docente, pela didática exemplar e pelo privilégio de compartilhar seu conhecimento e, como orientador, pela confiança, sensibilidade, paciência e orientação, além da oportunidade de desenvolver este estudo, sendo essencial não apenas para a minha formação acadêmica, mas também para o meu crescimento pessoal e profissional.

Agradeço à Universidade Federal de Ouro Preto e a todos os professores do Departamento de Estatística, pelo conhecimento compartilhado, que transformou minha jornada acadêmica, contribuiu significativamente para minha formação e despertou em mim um profundo interesse pela Estatística.

Por fim, agradeço aos amigos que fiz ao longo do curso, que tornaram esta caminhada mais leve, significativa e inesquecível.

A todos que, de alguma forma, estiveram presentes ao longo deste percurso, deixo o meu mais sincero agradecimento.

Resumo

Este estudo tem interesse em investigar as estratégias de otimização heurística aplicadas para redes de filas. É um problema instigante e desafiador no âmbito da investigação científica. Particularmente este estudo tem intuito de analisar a alocação de áreas de circulação (do inglês, *buffers*) em redes com filas de servidor único. É preponderante que a alocação de *buffers* seja planejada de forma a atender questões orçamentárias mas que seja capaz de satisfazer requisitos de eficiência em atendimento. Usualmente a eficiência em atendimento para problemas dessa natureza está associada com a taxa de atendimentos (do inglês, *throughput*). Em geral, os sistemas de filas tem disponibilidade orçamentária limitada, ou seja, existem restrições tanto para a quantidade de servidores quanto para o espaço total em áreas de espera. O custo total envolvido no processo é afetado sobremaneira por esses impactos financeiros, nesse estudo com fila de servidor único, o foco central reside na alocação dos buffers. O problema de alocação conjunta de áreas de espera, conhecido como BAP (do inglês *Buffer Allocation Problem*), é principal alvo desse estudo. É importante salientar que trata-se de um problema que não possui um tratamento simplista do ponto de vista computacional, isso se deve ao fato de se tratar de um problema de programação não linear, cuja solução analítica não apresenta forma fechada, consequentemente devem ser utilizados métodos aproximados em sua solução. Com interesse em incentivar e iniciar pesquisadores, este projeto visa estudar alguns conceitos da clássica Teorias das filas e também investigações de otimização combinatória, principalmente, a otimização em sistemas de filas. Estes sistemas de filas estão presentes em diversos cenários: o fluxo de tráfego em grandes centros urbanos, serviços telefônicos de atendimento (*call-center*), serviços de emergência médica, atendimentos de serviços públicos, entre outros. A metodologia empregada nesse trabalho pode ser utilizada para generalizar o estudo de diversos sistemas de filas em situações reais. As conclusões apresentadas, foram obtidas através da análise de variadas redes e podem auxiliar aos profissionais da área no planejamento de redes de filas gerais.

Palavras-chaves: Alocação de áreas de espera, filas finitas, redes de filas, otimização.

Abstract

This study is concerned with investigating heuristic optimization strategies applied to queueing networks. It is an intriguing and challenging problem within the scope of scientific research. In particular, this study aims to analyze the allocation of buffer areas in single-server queueing networks. It is essential that buffer allocation be planned in a way that addresses budgetary constraints while remaining capable of satisfying service efficiency requirements. Typically, service efficiency in problems of this nature is associated with the throughput. In general, queueing systems have limited budget availability; that is, there are constraints both on the number of servers and on the total buffers allocated areas. The total cost involved in the process is significantly affected by these financial impacts; in this single-server queueing study, the central focus lies on buffer allocation. The joint allocation problem of waiting areas, known as the Buffer Allocation Problem (BAP), is the main target of this study. It is important to emphasize that this problem does not admit a simplistic treatment from a computational standpoint, due to the fact that it is a nonlinear programming problem whose analytical solution does not have a closed form; consequently, approximate methods must be employed in its solution. With the aim of encouraging and initiating researchers, this project seeks to study some concepts from classical Queueing Theory as well as investigations in combinatorial optimization, primarily optimization in queueing systems. These queueing systems are present in various scenarios, such as traffic flow in large urban centers, telephone service centers (call centers), emergency medical services, and public service provision, among others. The methodology employed in this work can be used to generalize the study of various queueing systems in real-world situations. The conclusions presented were obtained through the analysis of several networks and may assist professionals in the field in the planning of general queueing networks.

Key-words: buffer allocation; finite queues; queueing networks; optimization.

Lista de figuras

Figura 1 – Rede de filas hipotética.	8
Figura 2 – Algoritmo NSGA-II	9
Figura 3 – Algoritmo Guloso	11
Figura 4 – Estruturas topológicas em investigação.	14
Figura 5 – Evolução do <i>throughput</i> entre as soluções investigadas.	15
Figura 6 – <i>Buffers</i> alocados na soluções apresentadas.	16
Figura 7 – Alocação de <i>buffers</i> nas filas paralelas.	17
Figura 8 – Alocação de <i>buffers</i> na fila inicial da rede.	18

Lista de tabelas

Tabela 1	–	Parâmetros em filas de acordo com a notação de KENDALL (1953) adaptada de SOUZA (2020)	6
Tabela 2	–	Performance média (<i>throughput</i>) do NSGA-II e do NSGA-II + Guloso.	16
Tabela 3	–	Alocação média de <i>buffers</i> do NSGA-II e do NSGA-II + Guloso. . . .	19

Sumário

1	INTRODUÇÃO	1
1.1	Objetivos	2
2	MATERIAIS E MÉTODOS	3
2.1	Otimização	3
2.2	Otimização em Redes de Filas	4
2.3	Teoria de Filas	5
2.3.1	Notação de Kendall	6
2.4	O Problema de Alocação de Áreas de Espera	6
2.5	Economia de Escala em Teoria das Filas	7
2.6	Algoritmo de Otimização Multi-objetivo	9
2.7	Algoritmo de Pós-processamento Guloso	10
3	RESULTADOS ALCANÇADOS	13
4	CONSIDERAÇÕES FINAIS	21
	REFERÊNCIAS	23

1 Introdução

Os procedimentos de otimização são fundamentais no dia a dia, são utilizados para minimizar custos e otimizar tarefas, por exemplo. A essência dessas ações é converter problemas práticos em formulações matemáticas para maximizar ou minimizar (otimizar) funções associadas com os objetivos relevantes nos problemas investigados. Em aplicações como sistemas de filas, a otimização pode aumentar significativamente a eficiência ao ajustar fatores como o número de servidores utilizados e alocação de espaços de espera ([YANG, 2010](#)).

Os métodos de otimização usualmente nominados exatos buscam soluções precisas e ótimas para problemas específicos. Usualmente essa busca ocorre por meio do uso de derivadas para progredir em um espaço de busca determinado e bem-definido. Esses métodos, exigem que a função objetivo e as restrições sejam contínuas e diferenciáveis, é importante salientar que estas são exigências demasiadamente restritivas. Embora ofereçam a garantia de encontrar a solução ótima global após um número finito de iterações, em problemas complexos ou de grande escala, o esforço computacional tende a ser também um problema. Em contraste, os métodos heurísticos não garantem a solução ótima global, mas são práticos para decisões cotidianas, como escolher rotas ou agendar compromissos, fornece soluções rápidas e frequentemente satisfatórias com um custo computacional significativamente menor ([MARTÍ; REINELT, 2022](#)).

Os métodos heurísticos oferecem soluções que não são as melhores, ou seja, otimalidade global, mas são boas o suficiente para muitas situações cotidianas, como decidir trajetos ou horários de compromissos. Em situações específicas, os métodos heurísticos são capazes de fornecer aproximação de grande qualidade para as soluções de otimalidade global. Muitas das vezes, a relação entre o custo da busca pela otimalidade global e o grau de eficácia e eficiência das soluções sub-ótimas produzidas por heurísticas faz com que os mecanismos heurísticos se tornem altamente atrativos. As estratégias heurísticas são mais rápidas que os métodos exatos e se aproximam das soluções ideais a depender da situação. Este estudo em particular, trata de problemas de otimização abordados por métodos heurísticos de otimização com foco em sua aplicação prática ([JUAN *et al.*, 2023](#)).

Este estudo apresenta o contexto de otimização para sistemas de filas em rede, aborda as situações comuns de espera em bancos, hospitais, tráfego e *call centers*. Quando a demanda por algum serviço supera a capacidade de atendimento, surgem filas, compostas por diversos elementos, como padrão de chegada de usuários, tempo de atendimento e capacidade de espera. O objetivo é otimizar a alocação de recursos nesses sistemas para melhorar sua eficiência operacional, por meio de algoritmos de otimização

adequados.

Essa investigação visa delinear o propósito específico de utilizar um a heurística gulosa acoplada com o clássico algoritmo genético em sua versão NSGA-II para resolver um problema de otimização específico em redes de filas com chegadas markovianas, atendimentos gerais, com servidor único e *buffers* finitos. A abordagem se concentra em aprimorar a eficiência operacional desses sistemas de filas $M/G/1/k$, na notação de [KENDALL \(1953\)](#). Aqui, $M/G/1/k$ representa um modelo matemático para descrever a chegada de usuários, o tempo de serviço, o número de servidores disponíveis e a capacidade máxima do sistema de filas, respectivamente. A estratégia heurística selecionada como uma ferramenta promissora para encontrar soluções próximas do ótimo global neste contexto específico, permitir assim melhorar a gestão de recursos e reduzir o tempo de espera dos usuários.

1.1 Objetivos

Os objetivos do presente estudo são:

- apresentação de uma revisão da bibliográfica na área de modelagem e otimização de redes de filas finitas;
- utilização dos algoritmos genético NSGA-II ([CRUZ *et al.*, 2012](#)) e guloso ([DUARTE; CRUZ; SOUZA, 2024](#)), como ferramenta acoplada de resolução do BAP;
- análise do BAP via experimentos computacionais, para identificação de padrões nas soluções ótimas, com interesse na verificação de efeitos de economia de escala.

Este texto é iniciado com o presente capítulo introdutório que aborda os aspectos de pesquisa discutidos durante o projeto, e também o delineamento prévio de objetivos. Posteriormente, um capítulo de Materiais e Métodos detalha a metodologia estatística que foi discutida e particularmente as formulações matemáticas para problemas de otimização utilizados ao longo do Projeto. O terceiro capítulo apresenta de forma mais definitiva todo o conjunto de resultados alcançados, detalha situações de utilização da metodologia apresentada e conjunto experimental abordado. Por fim, o último capítulo apresenta as conclusões alcançadas através dessa investigação e também propostas de continuidade desse estudo.

2 Materiais e Métodos

Os procedimentos de otimização são essenciais e bastante recorrentes na vida cotidiana atual. Desde minimizar custos em compras até encontrar a rota mais rápida no tráfego urbano. O interesse, mesmo que intuitivo, é transformar esses problemas em formulações matemáticas na direção de maximizar ou minimizar alguma função. Dentre as mais variadas aplicações, em sistemas de filas a eficiência pode ser aprimorada através da otimização de elementos inerentes ao modelo de filas como o número de servidores disponíveis a quantidade de espaços de espera, entre diversos outros.

Diante disso, a investigação científica de problemas dessa natureza é bastante difundida. Não somente difundida, essa linha de estudo científico é bastante instigante em decorrência de sua vasta aplicabilidade.

2.1 Otimização

Otimizar, em uma análise simplista, significa encontrar a melhor maneira de realizar alguma tarefa, na busca por resultados que possam ser considerados superiores. Por exemplo, ao fazer compras, busca-se reduzir o valor gasto e melhorar a qualidade ou quantidade dos produtos adquiridos. Da mesma forma, ao planejar uma rota, procura-se o caminho mais rápido, que evite congestionamentos e outros possíveis contratempos.

Do ponto de vista científico, muitos desses problemas precisam ser expressos por meio de formulações matemáticas. Em termos matemáticos, otimizar envolve sempre a ação de maximizar ou minimizar alguma função de interesse prático. De acordo com [YANG \(2010\)](#), os estudos de otimização abrangem uma ampla variedade de aplicações e contextos. Qualquer problema que envolva a busca por um certo nível de otimalidade se enquadra na categoria de problemas de otimização.

De acordo com trabalhos anteriores de destaque, como os de [DEB et al. \(2002\)](#), [RINGUEST \(2012\)](#), [COLLETTE; SIARRY \(2004\)](#), destacam-se alguns conceitos fundamentais para abordar problemas de otimização. A função objetivo é o ponto central do problema de otimização, representa a métrica cujo interesse é melhorar ou otimizar. Em geral é expressa como uma formulação matemática que relaciona informações inerentes ao sistema.

Essas informações inerentes ao sistema, usualmente são abordadas como os argumentos que influenciam diretamente a função objetivo. A tarefa de identificação desses argumentos é essencial para compreender como eles afetam o desempenho do sistema. Atrélado aos argumentos, define-se o espaço de busca ou espaço de variáveis

decisórias, refere-se ao conjunto de todas as possíveis combinações dos parâmetros do argumentos que impactam o sistema. Por exemplo, se o problema é definido por quatro variáveis decisórias (argumentos) parâmetros, o espaço de busca é quadrimensional, e representa todas as combinações possíveis desses argumentos. A delimitação do espaço de busca ocorre por meio de possíveis restrições. Essas restrições são as especificações do problema que limitam os valores viáveis para os argumentos e da função objetivo. No caso particular de problemas associados com filas e redes de filas, essas restrições podem ser relacionadas à capacidade dos servidores, capacidade de *buffers* ou requisitos de tempo de resposta, por exemplo. Do ponto de vista de função na concepção matemática, os argumentos determinam elementos do domínio da função objetivo. É importante também compreender o papel dos elementos da imagem dessa função, ou seja, os objetivos de otimização como abordado por [BERTSIMAS; TSITSIKLIS \(1997\)](#) e [RARDIN \(1998\)](#).

O espaço de objetivos é o conjunto de todos os valores possíveis da função objetivo associados aos elementos do espaço de busca, em suma, o espaço de objetivos corresponde à imagem da função objetivo. Compreender esse espaço contribui para a visualização de metas alcançáveis e desejáveis. Dentre elementos pertencentes ao espaço de objetivos estão as soluções ótimas locais e globais, uma solução ótima local é aquela que é a melhor dentro de uma região específica do espaço de busca. Por outro lado, uma solução ótima global é a melhor solução possível em todo o espaço de busca, independentemente de sua localização ([STOIBER; KROMOSER, 2021](#)).

Diante disso, a otimização é um conjunto de técnicas utilizadas para melhorar o desempenho de sistemas complexos, como redes de filas, processos industriais, logística e telecomunicações. Pode ser aplicada em diversos contextos e envolve uma variedade de métodos, que são escolhidos com base na natureza do problema e nos objetivos a serem alcançados. No caso específico das redes de filas, o foco principal está na melhoria de métricas como tempos de espera, taxas de atendimento e custos operacionais ([STOIBER; KROMOSER, 2021](#)).

2.2 Otimização em Redes de Filas

Ao mencionar Redes de Filas, o presente estudo tem interesse pelas filas presentes no cotidiano das pessoas, as filas de atendimento como serviços de saúde, congestionamentos de tráfego, produção em série ou até mesmo em sistemas virtuais como espera em sistemas do tipo *call centers* e *e-commerce*. De acordo com [ABENSUR \(2011\)](#), existe uma relação íntima direta da formação de filas com a capacidade de atendimento do servidor e a demanda a ser atendida. É bastante simples convergir para a suposição de que nas situações cuja demanda por determinado serviço ultrapassa à capacidade de servir haverá formação de filas.

Os sistemas de representação de filas usualmente são composto por cinco componentes básicos. O modelo de chegada de usuários que está ligado ao tempo entre chegadas sucessivas de usuários ao estabelecimento. O modelo de serviço que refere-se ao tempo gasto em cada atendimento, ou seja, um tempo próprio para cada execução de tarefa específica. Servidores ou canais disponíveis que indica o número de atendentes que efetuam simultaneamente os atendimentos. Os buffers ou capacidade representam o número máximo permitido de pessoas ou processos que aguardam atendimento. A disciplina da fila que são as regras que determinam a ordem em que os clientes serão atendidos. Essa representação está em consonância com a clássica notação de [KENDALL \(1953\)](#) que será detalhada posteriormente.

Nesse estudo, a concentração reside na resolução de problemas de otimização em sistemas de filas, com interesse em aprimorar a eficiência operacional dos sistemas. A otimização consiste no processo de buscar a configuração de utilização de recursos que seja mais eficaz para o sistema, seja para minimizar ou maximizar determinadas métricas de desempenho. Essas métricas podem envolver variáveis discretas ou contínuas, como o número de servidores, capacidade de *buffers*, entre outros ([HAVIV; RAVNER, 2021](#)).

Para resolver problemas de otimização em sistemas de filas, é essencial seguir uma abordagem sistemática, que envolve identificar claramente a função objetivo, os parâmetros relevantes, o espaço de busca e as restrições envolvidas. Além disso, a seleção e implementação de algoritmos de otimização adequados desempenham um papel crucial na busca por soluções ótimas. Ao aplicar esses conceitos e técnicas é possível melhorar significativamente a eficiência e o desempenho dos sistemas de filas.

2.3 Teoria de Filas

A Teoria das Filas emergiu como um campo teórico dentro dos estudos de Processos Estocásticos, mas rapidamente se expandiu para aplicações práticas em áreas como Estatística, Pesquisa Operacional e Matemática Aplicada. É um ramo dedicado a estudar sistemas em que a demanda por serviços pode superar a capacidade de atendimento disponível. Para isso, conta com uma linguagem própria e conceitos específicos, o que faz fundamental compreender termos como clientes, servidores, *buffers* e taxas de serviço, que ajudam a descrever e analisar o comportamento desses sistemas.

Na Teoria das Filas, os clientes representam qualquer entidade que necessita de atendimento, não se limitando a pessoas. Os servidores são os recursos que realizam o serviço, como atendentes, máquinas ou processadores e estão diretamente ligados ao desempenho e ao custo do sistema. Os *buffers*, ou filas de espera, correspondem ao espaço para os clientes que aguardam atendimento. Embora possam ser ilimitados em

alguns contextos, neste estudo eles são limitados e considerados um recurso importante. As taxas descrevem a frequência dos eventos: a taxa de chegada (λ) mede a entrada de clientes no sistema, enquanto a taxa de serviço (μ) indica a quantidade de atendimentos realizados por unidade de tempo.

2.3.1 Notação de Kendall

A notação mais difundida para representar modelos de filas é a extremamente conhecida notação de [KENDALL \(1953\)](#). As filas são descritas em um formato reduzido $A/B/m/k$ em que: A representa o modelo de distribuição de probabilidades que descreve o tempo entre chegadas; B representa o modelo de distribuição de probabilidades que descreve o tempo de serviço; m determina a quantidade de servidores disponíveis; e k determina a área de circulação, ou seja o espaço total disponibilizado para os clientes em espera (*buffers*) acrescidos do número de clientes que já estão em atendimento. A representação proposta por [KENDALL \(1953\)](#) fica melhor descrita nas representações da Tabela 1.

Tabela 1 – Parâmetros em filas de acordo com a notação de [KENDALL \(1953\)](#) adaptada de [SOUZA \(2020\)](#).

parâmetro	símbolo	definição
A e B	M	Exponencial
	D	Determinístico
	E_p	Erlang $_p$ ($p = 1, 2, \dots$)
	G	Geral
m	$1, 2, \dots$	número de servidores
k	$1, 2, \dots$	espaço total (servidores + área de espera)

As características de um sistema de filas são basicamente:

- i) padrão de chegadas dos clientes;
- ii) padrão de serviço dos canais de atendimento;
- iii) capacidade do sistema;
- iv) número de canais de serviço;
- v) número de estágios de serviço.

2.4 O Problema de Alocação de Áreas de Espera

Baseado nos problemas de filas, é preciso estabelecer uma formulação matemática adequada para os propósitos de investigação, a literatura já apresentou diversos

formatos: o Problema de Alocação de Servidores (SAP) discutido em DUARTE (2024), o Problema de Alocação de Áreas de Espera (BAP) abordado em uma formulação mono-objetivo por SMITH; CRUZ (2005), o Problema de Alocação de Áreas de Espera e de Servidores (BSAP) em uma formulação mono-objetivo discutido em OLIVEIRA *et al.* (2025), entre muitas outras. Nesse estudo particular, existe o interesse no BAP abordado em uma formulação multiobjetivo por CRUZ *et al.* (2012), CRUZ; DUARTE; SOUZA (2018) e DUARTE; CRUZ; SOUZA (2024). A formulação de programação matemática multiobjetivo mencionada é apresentada a seguir:

$$\min F(\mathbf{K}, \boldsymbol{\mu}) = \left[f_1(\mathbf{K}), f_2(\boldsymbol{\mu}), f_3(\mathbf{K}, \boldsymbol{\mu}) \right], \quad (2.1a)$$

sujeito a:

$$\begin{aligned} k_i &\in \mathbb{N}, \forall i \in \{1, 2, \dots, m\}, \\ \mu_i &\geq 0, \forall i \in \{1, 2, \dots, m\}, \end{aligned} \quad (2.1b)$$

que compreende a minimização das capacidades e taxas de serviço, simultaneamente à maximização da taxa de atendimento (*throughput*). Então, $f_1(\mathbf{K}) = \sum_{i=1}^m k_i$ representa as capacidades totais (*buffers* acrescidos de servidores), $f_2(\boldsymbol{\mu}) = \sum_{i=1}^m \mu_i$ representa as taxas de serviço gerais e $f_3(\mathbf{K}, \boldsymbol{\mu}) = -\Theta(\mathbf{K}, \boldsymbol{\mu})$ representa o *throughput*. O sinal de menos no *throughput* se deve ao fato de que esse objetivo deve ser maximizado. Uma abordagem para calcular o *throughput* em redes de filas $M/G/1/k$ é necessária para usar essa formulação.

Essa última formulação é de fato o interesse direto desse estudo. Em outras palavras, o objetivo é minimizar o consumo de recursos mas garantir o atendimento de algum limiar mínimo de taxa de atendimento. Como essa formulação já foi abordada em estudo anterior, o interesse aqui está em constatar se a estratégia gulosa de otimização alcança melhorias por meio da economia de escala em alocação de *buffers*.

2.5 Economia de Escala em Teoria das Filas

O conceito de economia de escala é amplamente aplicado em diversas áreas da atividade econômica, como a indústria, agricultura, tecnologia, varejo, serviços públicos, saúde e educação. A economia de escala ocorre quando há um aumento na escala de produção de um determinado produto ou serviço, por um certo período, e esse aumento resulta na redução do custo unitário desse produto ou serviço. Quando isso ocorre, temos o que se chama de economia de escala como mencionado por CHAOS *et al.* (2021).

Dentro desse conceito, há uma subdivisão em duas categorias: a economia de escala interna, que ocorre dentro da empresa e está relacionada à capacidade da organização de otimizar a produção por meio de melhorias técnicas, e a economia

de escala externa, que ocorre fora da organização, geralmente como resultado do crescimento do setor ou da concentração de empresas em uma mesma região, o que gera benefícios compartilhados.

A Teoria das filas é uma área da Estatística que estuda os fenômenos relacionados à espera de clientes por algum serviço. Em sistemas com diversas filas interconectadas a economia de escala pode ser um relevante efeito alcançado por meio de estratégias de otimização. Esse é particularmente um foco desse estudo.

A Figura 1 pode auxiliar na compreensão desse efeito, é apresentada uma rede com m filas, existe uma fila de entrada na rede e depois uma divisão para $m - 1$ filas subsequentes. Nessa estrutura topológica, a primeira fila da rede é claramente um gargalo que pode apresentar impactos significativos no desempenho da rede. O interesse é verificar se o algoritmo guloso dessa investigação alcança melhora nas taxas de atendimento por meio de efeito de economia de escala em filas paralelas para consumir os *buffers* economizados em filas com característica de gargalos. O objetivo é verificar se a estratégia gulosa desse estudo propões soluções que na analogia da Figura 1 teriam uma redução em $\sum_{i=2}^m k_i$ para alocar esses *buffers* economizados na fila inicial da rede e assim alcançar melhora no *throughput* Θ da rede de filas.

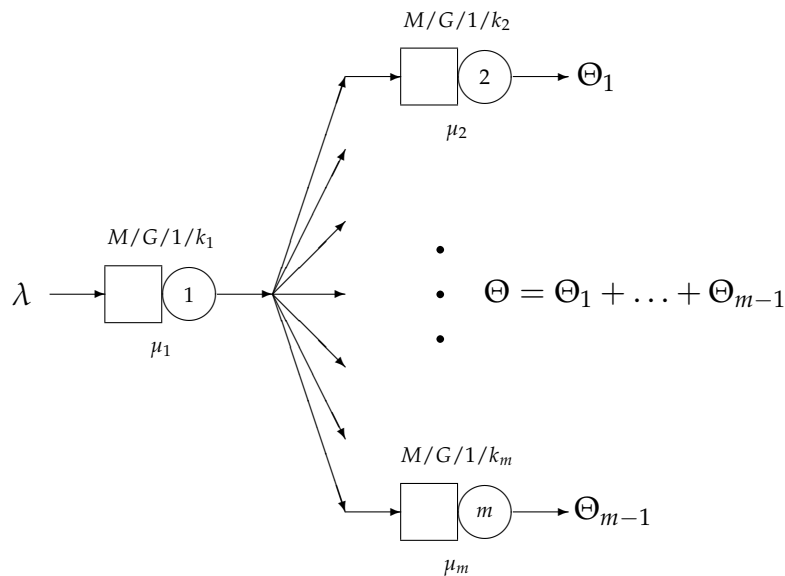


Figura 1 – Rede de filas hipotética.

O efeito de economia de escala permite alcançar melhorias de desempenho sem que ocorra demanda por novos investimentos. Esse tipo de análise é útil em diversos contextos de atendimento de clientes (CARLSSON; PENG; RYZHOV, 2024). A economia de escala no contexto da Teoria das filas refere-se à diminuição /realocação dos custos médios de operação à medida que a performance de um sistema de filas aumenta, ou seja, reduzir ou realocar o número de servidores ou a capacidade de atendimento,

o sistema pode se tornar mais eficiente em termos de custo por unidade de serviço prestado (GHAZALI; TAHAR, 2025).

Em sistemas de filas, como em muitos outros setores, a economia de escala surge quando o recurso consumido é economizado sem perda de performance ou até mesmo com ganho de performance. Isso ocorre porque, ao utilizar de forma eficaz o número de servidores ou a infraestrutura, isso pode acarretar em reduções no tempo de espera ou tempo de atendimento dos clientes, o que conduz para a redução de congestionamento e custo operacional.

2.6 Algoritmo de Otimização Multi-objetivo

A adaptação de um algoritmo evolucionário para fornecer soluções ao problema de otimização descrito pelas Equações (2.1a) e (2.1b) foi abordada por CRUZ *et al.* (2012). Para um algoritmo evolucionário multiobjetivo (MOEA), a aplicação dos operadores genéticos *mutação*, *cruzamento*, *seleção* e *elitismo* garante a convergência da população para uma frente de Pareto. O MOEA utilizado é o algoritmo genético *Non-dominated Sorting Genetic Algorithm II* (NSGA-II) (DEB *et al.*, 2002), implementado especificamente para o problema de otimização de redes de filas. Os detalhes de configuração específicos para o NSGA-II usado nesse estudo foram devidamente calibrados por CRUZ *et al.* (2012). Outras informações também podem ser obtidas em CRUZ; DUARTE; SOUZA (2018) (consulte Algoritmo da Figura 2).

```

algoritmo
   $P_1 \leftarrow \text{GeraPopulaçãoInicial}(\text{popSize})$ 
  para  $i = 1$  até numGen faça
    /* gera filhos por cruzamento e mutação */
     $Q_i \leftarrow \text{FaçaPopulaçãoNova}(P_i)$ 
    /* combina pais e filhos */
     $R_i \leftarrow P_i \cup Q_i$ 
    /* encontre fronteiras não-dominadas  $\mathcal{F} = (\mathcal{F}_1, \mathcal{F}_2, \dots)$  */
     $\mathcal{F} \leftarrow \text{OrdenaçãoNãoDominante}(R_i)$ 
    /* encontre nova população pela distância de aglomeração */
     $P_{i+1} \leftarrow \text{GeraPopulaçãoNova}(R_i)$ 
  fim para
   $P_{\text{numGen}+1} \leftarrow \text{ExtraiPareto}(P_{\text{numGen}})$ 
  escreva  $P_{\text{numGen}+1}$ 
fim algoritmo

```

Figura 2 – Algoritmo NSGA-II

2.7 Algoritmo de Pós-processamento Guloso

O NSGA-II discutido anteriormente produz um conjunto de Pareto subótimo. Não há garantia de que todas as combinações do vetor de alocação de *buffers* sejam avaliadas. Os operadores genéticos não buscam todas as alocações de *buffers* possíveis que preservem a mesma capacidade total, $\sum_{i=1}^m k_i$, e forneçam a melhor taxa de atendimento possível Θ . Alguma recombinação de *buffers* no sistema de filas pode levar a alguma melhoria na taxa de atendimento. Por muitas gerações, o algoritmo NSGA-II tende a aproximar (em termos de convergência) a melhor solução que o método pode obter. Nesse ponto, um número adicional proibitivo de gerações é necessário para pequenas etapas de melhoria.

CRUZ; DUARTE; SOUZA (2018) observaram que melhorias poderiam ser obtidas com pequenas alterações no vetor de decisão $\mathbf{k}$ e avaliou esse fato por meio de uma tentativa de pós-processamento pelo algoritmo *Simulated Annealing*. Posteriormente, DUARTE; CRUZ; SOUZA (2024) verificaram que o pós-processamento pode ser executado por uma estratégia computacionalmente mais econômica, e ainda alcançar resultados promissores. O algoritmo guloso é usado para a nova abordagem.

Estratégias gulosas frequentemente não garantem a otimalidade, mas, quando aplicáveis, são mais eficientes do que outras técnicas, como programação dinâmica ou busca exaustiva. Este estudo adota uma estratégia gulosa que recebe uma solução inicial NSGA-II. Dado o valor $\sum_{i=1}^m k_i$ da solução anterior, o algoritmo guloso parte de uma solução proposta com um único espaço de *buffer* para cada fila na rede. Posteriormente, o algoritmo escolhe qual fila na rede receberá mais um espaço de *buffer* e proporcionará o ganho mais significativo possível na melhoria de *throughput*. Enquanto o *throughput* é melhorado, o procedimento é repetido desde que atenda ao critério de parada de que a alocação total de *buffers* seja menor ou igual à alocação prévia fornecida pelo NSGA-II.

A aplicação deste algoritmo guloso garante o fornecimento de soluções com um custo de alocação de *buffers* menor ou igual ao das soluções anteriores do NSGA-II. Por outro lado, as soluções fornecidas têm um *throughput* igual ou superior à das soluções anteriores. Além disso, este procedimento é realizado com baixo custo computacional, o que torna a técnica a proposta bastante atrativa.

O principal benefício alcançados pela otimização por meio do algoritmo guloso é a redução do espaço de *buffers* alocados e ainda assim ampliação da taxa de atendimento. O presente estudo busca verificar se este efeito é alcançado pelo tratamento de filas gargalo por meio de economia de escala. Uma versão do pseudo-código para o *fine-tuning* do algoritmo guloso para o problema em estudo é apresentado na Figura 3.

algoritmo

```

sol ← RecebeSoluçõesNSGA-II(popSize)
/* soli é a i-ésima solução do NSGA-II */
para i = 1 até popSize faça
    Sk ← k1 + k2 + ... + km
    θk,μ ← Θ(k1, k2, ..., km, μ1, μ2, ..., μm)
    para j = 1 até m faça
        ki* ← 1
    fim para
    j ← m
    θk,μ* ← Θ(k1*, k2*, ..., km*, μ1, μ2, ..., μm)
    θmax ← θk,μ*
    enquanto j < Sk faça
        ki* ← 1
        para q = 1 até m faça
            ki* ← ki* + 1
            θk,μ* ← Θ(k1*, k2*, ..., km*, μ1, μ2, ..., μm)
            se θk,μ* > θmax então
                θmax ← θk,μ*
                ℓ ← q
            fim se
            ki* ← ki* - 1
        fim para
        Sk* ← k1* + k2* + ... + km* + 1
        se ℓ > 0 e Sk* ≤ Sk então kℓ* ← kℓ* + 1 fim se
        senão j ← Sk
        j ← j + 1
    fim enquanto
    θk,μ* ← Θ(k1*, k2*, ..., km*, μ1, μ2, ..., μm)
    se θk,μ* ≥ θk,μ então soli ← (k1*, k2*, ..., km*, μ1, μ2, ..., μm) fim se
    senão soli ← (k1, k2, ..., km, μ1, μ2, ..., μm)
fim para
fim algoritmo

```

Figura 3 – Algoritmo Guloso

A seção subsequente discute a utilização dos dois algoritmos, o NSGA-II puro e

depois acoplado ao algoritmo guloso com interesse e avaliar os efeitos resultantes de suas estratégias de otimização. Toda a investigação é conduzida com o propósito de verificar qual a estratégia o algoritmo guloso utiliza para melhorar os resultados.

3 Resultados Alcançados

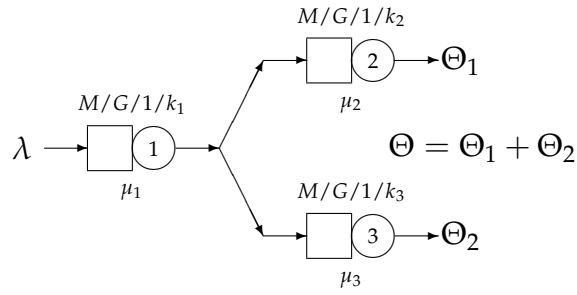
Os algoritmos de otimização do presente estudo foram implementados em FORTRAN para compatibilidade com implementações anteriores do NSGA-II (CRUZ *et al.*, 2012) e do GEM apresentadas em SMITH (2003) e SMITH (2004). Todos os códigos estão disponíveis mediante solicitação para fins educacionais e de pesquisa. Os experimentos computacionais foram conduzidos no processador Intel® Core™ i7-1165G7 de 11ª geração de 2,80 GHz rodando Windows 11 Home de 64 bits, com 8,00 GB de RAM.

Topologias variadas de redes de filas foram utilizadas na condução do experimento desse estudo. O objetivo foi o de avaliar a adaptação do algoritmo proposto em diversos cenários. Sem perda de generalidade, todos os experimentos consideraram servidores com a mesma capacidade de trabalho, ou seja, a mesma taxa de serviço μ . No entanto, é relevante citar que não existem quaisquer restrições acerca da realização de novos experimentos com servidores de capacidades distintas e outras possibilidades experimentais.

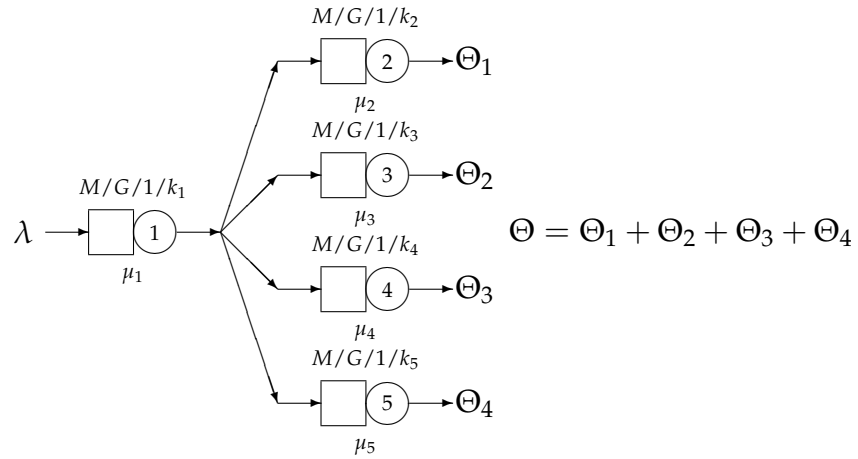
De acordo com estudos relatados anteriormente por CRUZ *et al.* (2012), o melhor grupo de parâmetros para NSGA-II é o uso combinado de cruzamento binário simulado (SBX) e de mutação, com uma taxa de mutação de 2%, 400 indivíduos pareciam ser suficientes, o parâmetro de dispersão deveria ser aproximadamente 8, e foi definido o número máximo de gerações numGen para 4.000 (isso garantia um tempo de computação finito).

O principal objetivo do pós-processamento guloso é refazer a alocação de *buffers* na rede de filas. As novas soluções fornecidas são iguais ou superiores às soluções anteriormente fornecidas pelo NSGA-II. O algoritmo deve ser capaz de reduzir o custo em recursos (*buffers*) e ainda garantir a preservação ou aumento do *throughput* da rede de filas.

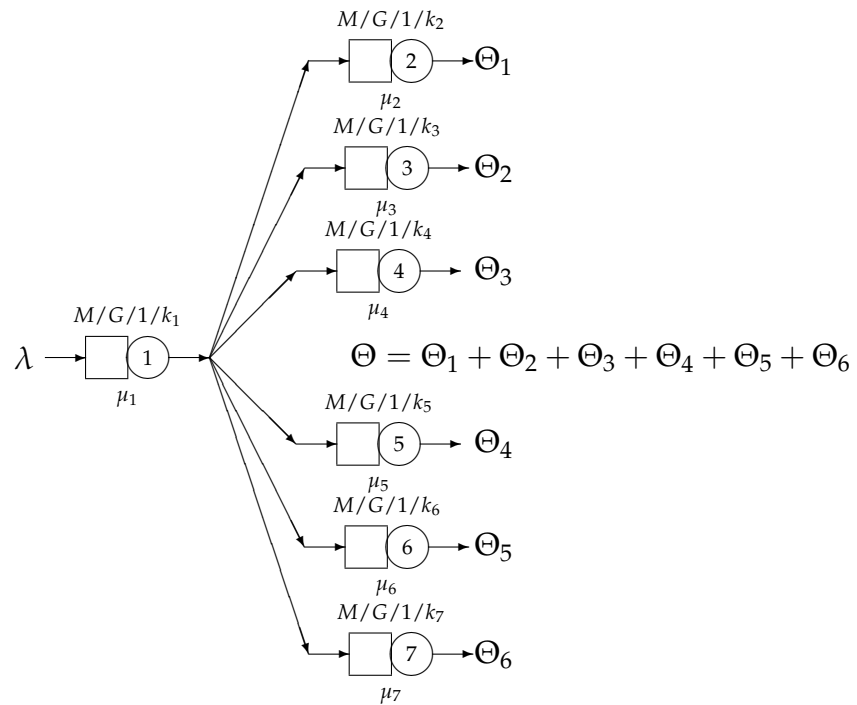
É importante ainda, observar que essa formulação visa garantir que a redução orçamentaria para o funcionamento da rede de filas não resulte em uma perda da qualidade do desempenho. Essa perda de desempenho poderia ser aferida em diversas métricas, particularmente aqui é mensurada em termos do *throughput* da rede de filas. Os algoritmos são capazes de otimizar quaisquer topologias, entretanto o objetivo particular aqui é identificar possíveis efeitos de melhora via economia de escala em redes de topologia divisão apresentadas na Figura 4.



(a) Rede de filas em divisão com 3 nós.



(b) Rede de filas em divisão com 5 nós.



(c) Rede de filas em divisão com 7 nós.

Figura 4 – Estruturas topológicas em investigação.

A Figura 5 apresenta as medidas de *throughput* alcançadas para todas as redes de filas em estudo (ver Figura 4). A caracterização da distribuição geral foi inserida por meio de variações no quadrado do coeficiente de variação dos tempos de serviço. Dessa forma, foram utilizados serviços hipoexponenciais ($cv^2 = 0,5$), Markovianos ($cv^2 = 1,0$), e ainda hiperexponenciais ($cv^2 = 1,5$).

Para todas as variações, os experimentos foram conduzidos com o algoritmo NSGA-II puro e o algoritmo NSGA-II combinado com a heurística gulosa. Estão consideradas as 400 soluções finais fornecidas. Os resultados mostram que a inclusão da heurística gulosa proporcionou ganhos notórios de desempenho evidentes em *throughput* como já havia sido relatado por DUARTE; CRUZ; SOUZA (2024).

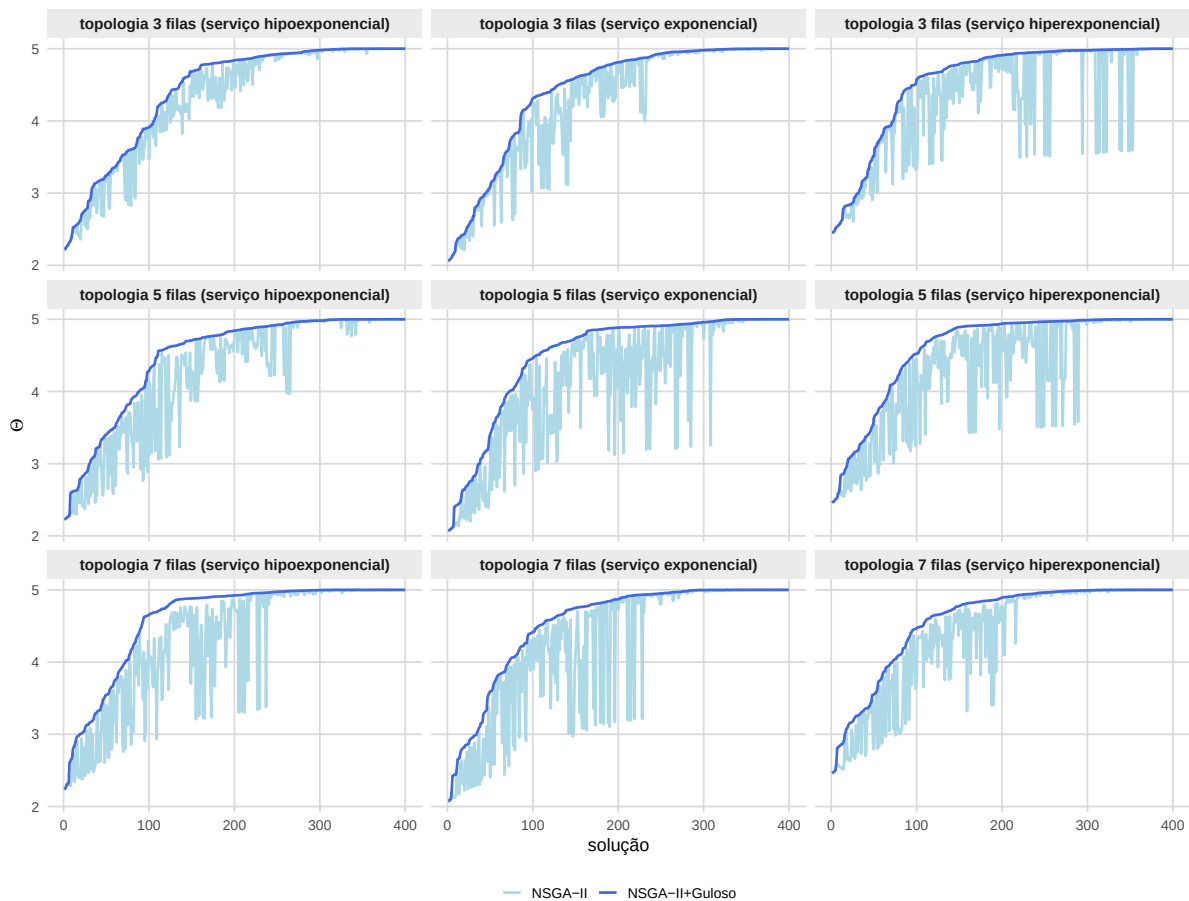


Figura 5 – Evolução do *throughput* entre as soluções investigadas.

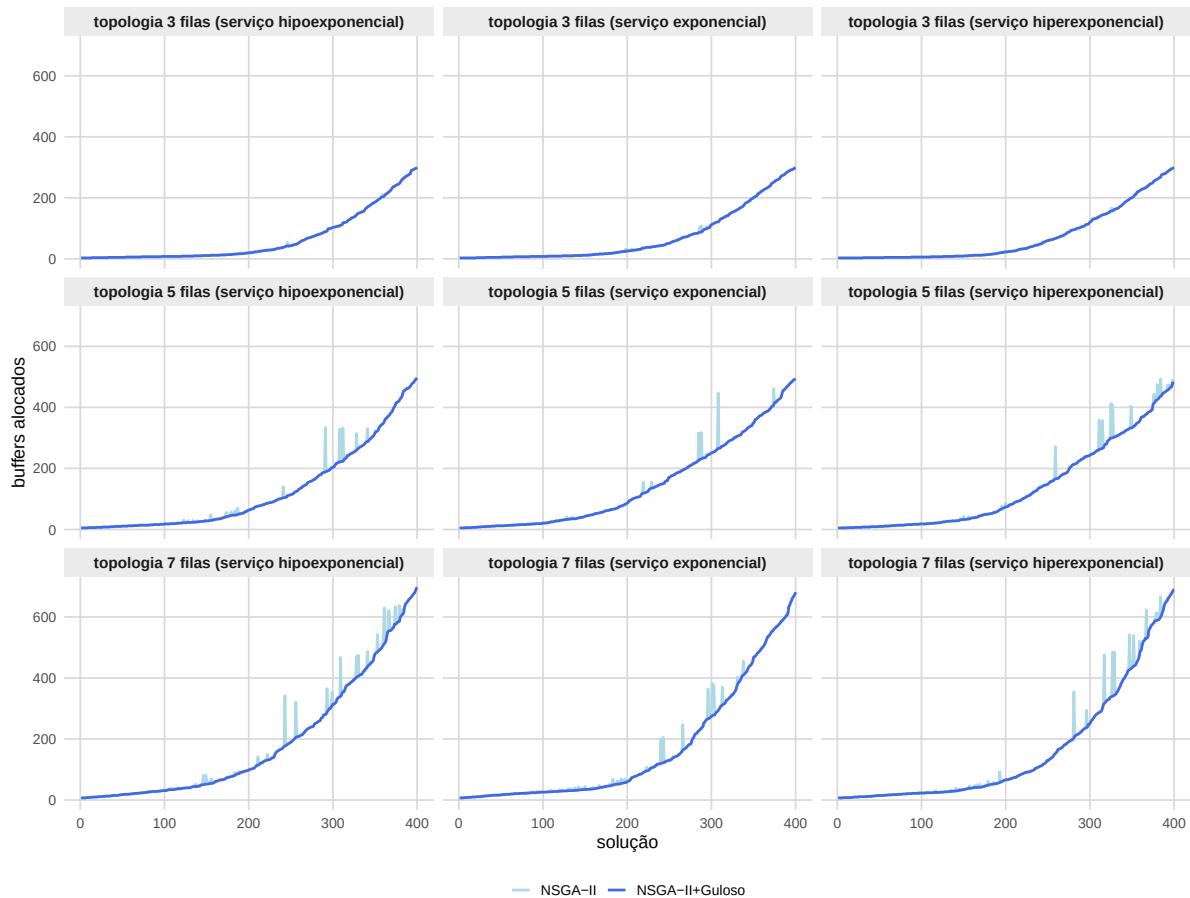
As constatações da Figura 5 são confirmadas em uma análise quatitativa por meio dos efeitos de locação e escala. Essa avaliação é apresentada na Tabela 2, independente da topologia e da distribuição do serviço, as soluções fornecidas pela combinação do NSGA-II com o algoritmo Guloso fornecem em média *throughput* superior e com menor variabilidade.

Tabela 2 – Performance média (*throughput*) do NSGA-II e do NSGA-II + Guloso.

Topologia	cv^2	NSGA-II		NSGA-II + Guloso	
		média	d.p.	média	d.p.
3 filas	0,5	4,34	0,76	4,56	0,69
	1,0	4,30	0,83	4,42	0,78
	1,5	4,26	0,87	4,41	0,82
5 filas	0,5	4,33	0,77	4,60	0,66
	1,0	4,25	0,86	4,48	0,73
	1,5	4,15	0,89	4,50	0,76
7 filas	0,5	4,37	0,79	4,57	0,64
	1,0	4,30	0,86	4,59	0,68
	1,5	4,26	0,92	4,53	0,72

Nota: melhores resultados em **destaque** e d.p. representa o desvio-padrão entre as soluções.

O ganho constatado na Figura 5 e na Tabela 2 já era conhecido por meio dos estudos apresentados por DUARTE; CRUZ; SOUZA (2024). A pergunta específica está em discutir se esse ganho decorre somente de realocação de *buffers*, ou além disso existe economia real de *buffers*? A Figura 6 discute esse ponto.

Figura 6 – *Buffers* alocados na soluções apresentadas.

A medida que a dimensão da rede de filas, o algoritmo começa a ser capaz de

realizar economia real de *buffers*. Essa constatação já era previsível, com o aumento do número de filas também é ampliada a possibilidade de maiores diferenças baseadas na realocação de *buffers*.

Entretanto, o objetivo central desse estudo é verificar qual a estratégia que o algoritmo Guloso utiliza para alcançar vantagem no aumento de *throughput* sem ampliar o consumo de recursos em *buffers*. A principal suspeita decorre da possibilidade de efeito de economia de escala nas filas em paralelo do rede para utilização de áreas de *buffers* economizadas para serem alocadas em filas gargalos da rede. A Figura 7 apresenta a avaliações dos efeitos de soma de *buffers* alocados nas filas em paralelo para as redes de filas em estudo detalhadas na Figura 4.

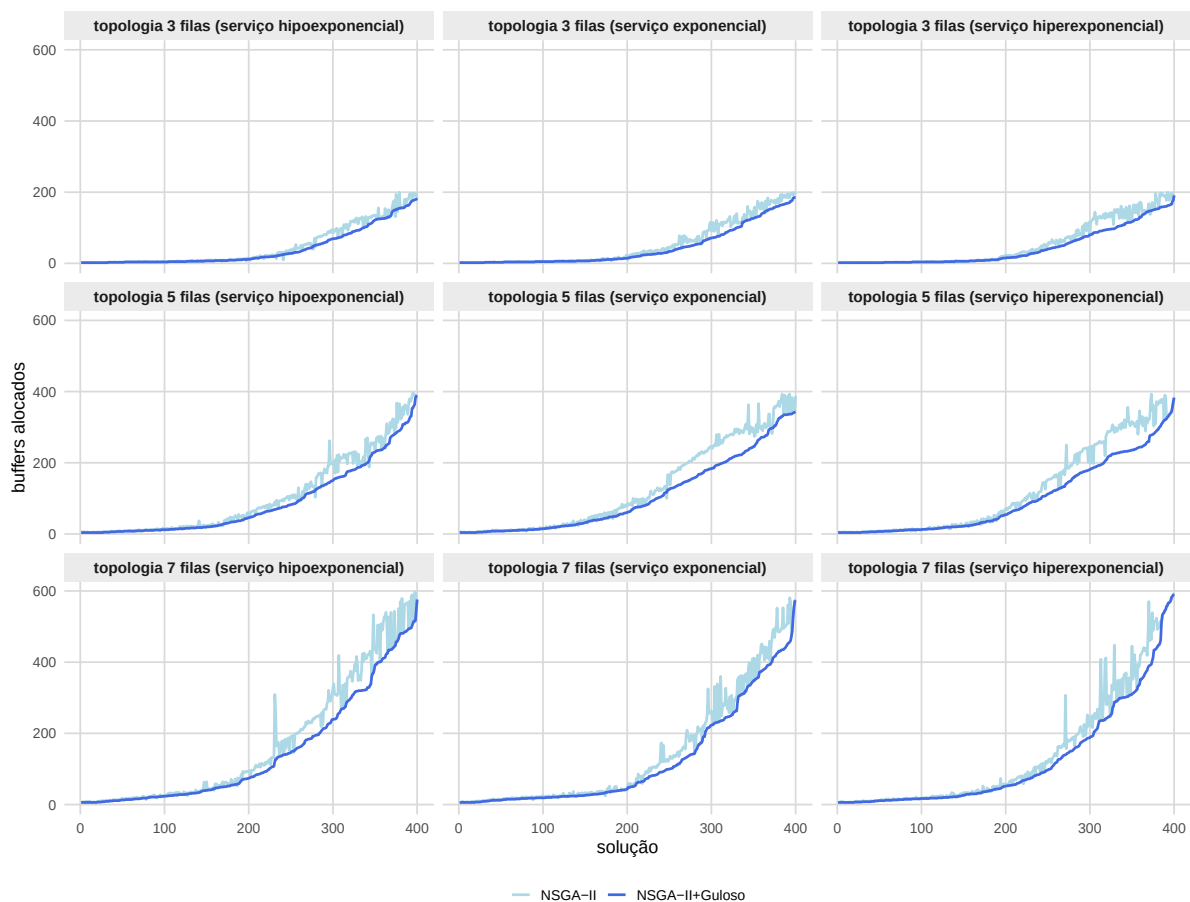


Figura 7 – Alocação de *buffers* nas filas paralelas.

A comparação apresenta o efeito de economia de escala executado nas filas em paralelo das soluções fornecidas pelo algoritmo acoplado NSGA-II + Guloso em comparação com as soluções provenientes do algoritmo NSGA-II puro. Para a rede com 3 filas, somaram-se as alocações das filas 2 e 3; para a rede com 5 filas, somaram-se as alocações das filas 2, 3, 4 e 5; e por fim, para a rede com 7 filas, somaram-se as alocações das filas 2, 3, 4, 5, 6 e 7.

Os efeitos que eram presumidos previamente se confirmam, de fato as soluções

do algoritmo acoplado NSGA-II + Guloso alocam sistematicamente menor soma de *buffers* nas filas em paralelo. Apesar dos gráficos da Figura 7 sugerirem que este efeito é ampliado a medida que a rede aumenta, mas esse efeito é ilusório.

É um efeito decorrente de uma soma total maior em redes maiores, uma rápida análise nos coeficientes de variação entre desvio-padrão e soma média de *buffers* desmistifica esse efeito visual. Em contraponto, a Figura 8 apresenta os efeitos na fila inicial que funciona como gargalo para o funcionamento da rede.

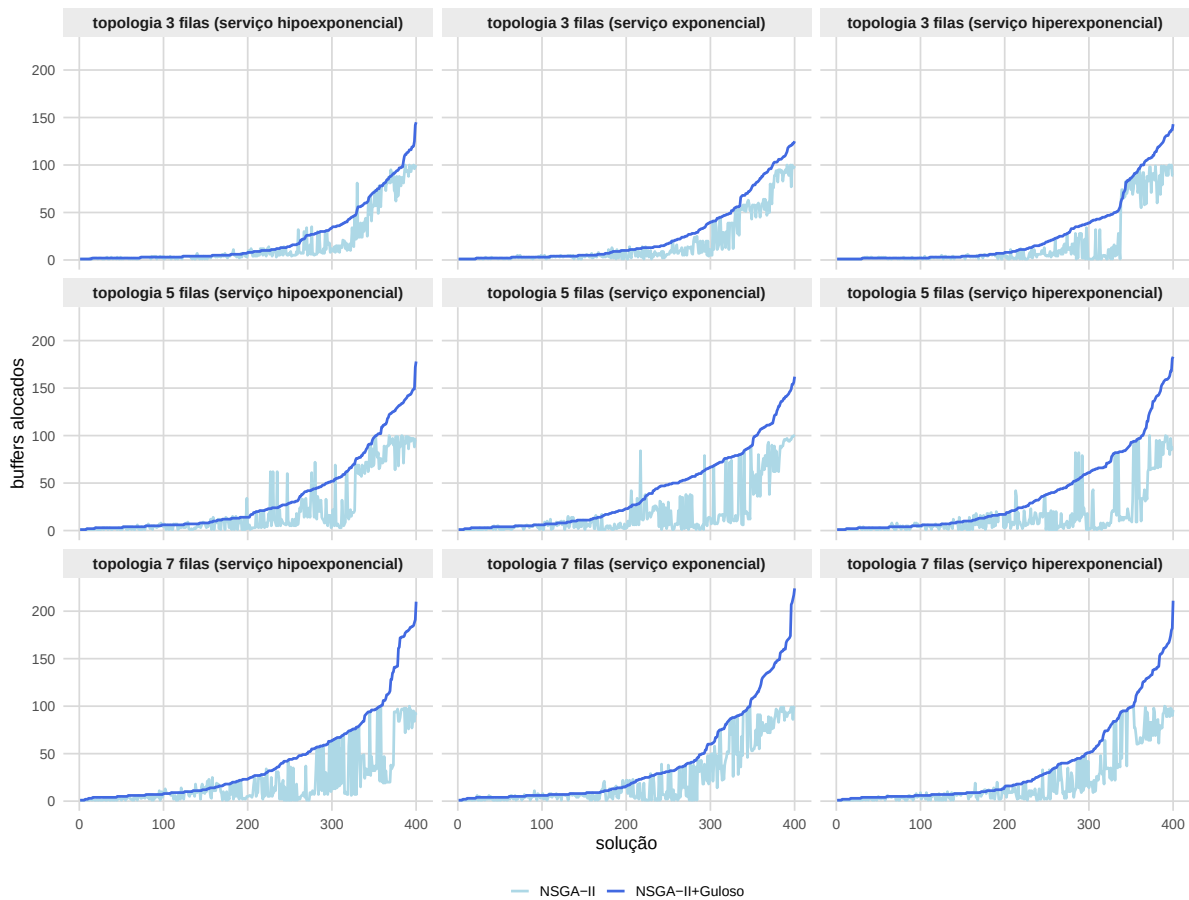


Figura 8 – Alocação de *buffers* na fila inicial da rede.

A análise da alocação de *buffers* na primeira fila da rede confirma a expectativa prévia. De fato a estratégia gulosa retira recursos das filas em paralelo e consome este recursos na fila inicial que funciona como um gargalo para a rede. Esse efeito notoriamente possibilita a ampliação do *throughput* alcançado nas soluções fornecidas pelo algoritmo Guloso combinado com o NSGA-II. A Tabela 3 apresenta uma análise quantitativa e comparativa entre as duas abordagens de otimização para os efeitos de realocação.

Tabela 3 – Alocação média de *buffers* do NSGA-II e do NSGA-II + Guloso.

Topologia	cv^2	NSGA-II				NSGA-II + Guloso		
		fila paralelas	fila 1	todas as filas	fila paralelas	fila 1	todas as filas	
3 filas	0,5	54,1 (62,5)	17,0 (29,4)	71,1 (82,9)	43,7 (51,5)	27,4 (37,1)	71,1 (80,6)	
	1,0	47,4 (57,2)	17,7 (27,3)	65,1 (76,5)	40,8 (50,6)	24,3 (32,1)	65,1 (74,7)	
	1,5	52,4 (61,0)	18,0 (26,2)	70,4 (79,2)	43,8 (53,0)	26,4 (33,4)	70,2 (78,4)	
5 filas	0,5	123,9 (125,1)	16,1 (24,6)	140,0 (135,7)	99,5 (100,8)	38,3 (43,2)	137,8 (130,0)	
	1,0	104,7 (110,2)	22,5 (30,9)	127,2 (127,1)	89,4 (97,1)	36,0 (41,7)	125,4 (125,8)	
	1,5	128,9 (122,7)	19,0 (27,1)	147,9 (135,8)	105,8 (103,3)	40,8 (40,8)	146,6 (130,1)	
7 filas	0,5	143,7 (165,9)	21,8 (29,4)	165,5 (176,3)	124,4 (148,3)	40,8 (40,8)	165,2 (161,1)	
	1,0	172,7 (176,2)	20,2 (26,5)	192,9 (182,7)	146,2 (153,0)	43,1 (46,9)	189,3 (180,9)	
	1,5	144,5 (161,4)	23,9 (30,8)	168,4 (173,3)	124,9 (142,9)	40,8 (49,3)	165,7 (173,2)	

Nota: melhores resultados em **destaque** e desvio-padrão entre parenteses.

Os resultados apresentados na Tabela 3 atestam que o algoritmo guloso alcança alguma redução no consumo médio de *buffers*. Entretanto deve-se salientar que trata-se de um ganho marginal. Que contrasta com os resultados da Tabela 2 que mostram um ganho médio em *throughput* representativo. Isso confirma a perspectiva prévia que o algoritmo guloso não alcançava uma melhoria de resultados aliada à redução efetiva de consumo de *buffers*. Por outro lado, o algoritmo guloso se mostrou capaz de alocar de forma mais adequada todo o recurso disponível e assim alcançar melhorias de desempenho sem ampliação do comprometimento orçamentário.

4 Considerações Finais

Durante a execução deste trabalho, foi possível atender aos objetivos previamente traçados. Inicialmente, foram executados levantamentos e leituras de diversos trabalhos já publicados sobre modelagem e otimização em redes de filas finitas, o que proporcionou o suporte teórico para o desenvolvimento do estudo. Em seguida foram explorados métodos clássicos de estimação de medidas de desempenho em filas finitas, o que consolidou a compreensão dos conceitos essenciais e serviu de base para as análises realizadas. A aplicação do algoritmo genético NSGA-II, em conjunto com a estratégia gulosa em pós-processamento, permitiu abordar de forma prática o Problema de Alocação de Buffers (BAP). Isso destacou o potencial das heurísticas multiobjetivo. Experimentos computacionais com diversas simulações foram conduzidas em distintas configurações de redes, o que possibilitou observar o comportamento em diferentes cenários. Esses experimentos revelaram padrões consistentes nas soluções de melhor desempenho e confirmaram a ocorrência de estratégia de economia de escala para que a heurística gulosa fosse capaz de fornecer soluções mais eficientes. Isso reforçou tanto a eficácia quanto a relevância da abordagem proposta.

O estudo fundamentou-se na Teoria das Filas e em conceitos de otimização multiobjetivo. A estratégia utilizou uma implementação do NSGA-II e da heurística gulosa em FORTRAN gentilmente cedida pelos autores originais da abordagem. O pós-processamento guloso se mostrou de fato eficiente em refinar soluções genéticas preliminares. Os experimentos computacionais foram conduzidos em diferentes topologias de redes de filas com divisão, compostas por 3, 5 e 7 nós, e para distintos padrões de variabilidade dos tempos de serviço. As variações no quadrado do coeficiente de variação forneceram estratégias de tempo de serviço com distribuições hipoexponenciais, exponenciais e hiperexponenciais. Para viabilizar essa análise, foram utilizados recursos computacionais adequados, que possibilitaram a avaliação sistemática do desempenho dos algoritmos em múltiplos cenários.

Os resultados mostraram que a integração entre o NSGA-II e o algoritmo guloso promove ganhos expressivos em termos de desempenho, amplia o *throughput* das redes de filas sem elevar o consumo de recursos. Ficou evidente que a estratégia gulosa redistribui *buffers* das filas paralelas para filas gargalo, o que melhora a eficiência da rede como um todo. Embora a economia real de *buffers* tenha se mostrado marginal, a melhoria obtida na utilização dos recursos disponíveis foi consistente. Esse resultado demonstrou a capacidade da abordagem em alcançar soluções mais eficazes sob restrições orçamentárias. Assim, o estudo evidencia que o método proposto pode contribuir para a otimização de sistemas práticos, com equilíbrio entre custos e performance.

Por outro lado, é importante destacar algumas limitações. O estudo restringiu-se a redes com servidores com a mesma capacidade de atendimento e topologias específicas, sem avançar para cenários com servidores com capacidade de atendimento distintas. Além disso, as análises foram realizadas em ambiente experimental controlado, não incorporando variáveis adicionais presentes em sistemas reais, como custos de implementação ou restrições operacionais mais amplas. Para trabalhos futuros, é importante ampliar os experimentos para redes com servidores de diferentes capacidades, explorar topologias mais sofisticadas e aplicar a metodologia em áreas práticas como tráfego urbano e sistemas de saúde entre outros. Tais desdobramentos podem fortalecer ainda mais a relevância e aplicabilidade das contribuições apresentadas neste estudo.

Referências

- ABENSUR, E. O. Banking operations using queuing theory and genetic algorithms. **Produto & Produção**, v. 12, n. 2, 2011. Citado na página 4.
- BERTSIMAS, D.; TSITSIKLIS, J. N. **Introduction to linear optimization**. [S.l.]: Athena scientific Belmont, MA, 1997. v. 6. Citado na página 4.
- CARLSSON, J. G.; PENG, X.; RYZHOV, I. O. Demand equilibria in spatial service systems. **Manufacturing & Service Operations Management**, INFORMS, v. 26, n. 6, p. 2305–2321, 2024. Citado na página 8.
- CHAOS, S. R. *et al.* Economies of scale in cruise shipping. **Maritime Economics & Logistics**, Springer, v. 23, n. 4, p. 674–696, 2021. Citado na página 7.
- COLLETTE, Y.; SIARRY, P. Multiobjective optimization: principles and case studies. **OR/MS Today**, Institute for Operations Research and the Management Sciences, v. 31, n. 1, p. 60–61, 2004. Citado na página 3.
- CRUZ, F. R. B.; DUARTE, A. R.; SOUZA, G. L. Multi-objective performance improvements of general finite single-server queueing networks. **Journal of Heuristics**, Springer, v. 24, n. 5, p. 757–781, 2018. Citado 3 vezes nas páginas 7, 9 e 10.
- CRUZ, F. R. B. *et al.* Throughput maximization of queueing networks with simultaneous minimization of service rates and buffers. **Mathematical Problems in Engineering**, Hindawi, v. 2012, n. Article ID 692593, p. 19 pages, 2012. Citado 4 vezes nas páginas 2, 7, 9 e 13.
- DEB, K. *et al.* A fast and elitist multiobjective genetic algorithm: NSGA-II. **IEEE Transactions on Evolutionary Computation**, IEEE, v. 6, n. 2, p. 182–197, 2002. Citado 2 vezes nas páginas 3 e 9.
- DUARTE, A. R. The server allocation problem for Markovian queueing networks. **International Journal of Services and Operations Management**, v. 48, n. 2, p. 256–271, 2024. Citado na página 7.
- DUARTE, A. R.; CRUZ, F. R. B.; SOUZA, G. L. A greedy post-processing strategy for multi-objective performance optimization of general single-server finite queueing networks. **Soft Computing**, Springer, v. 28, n. 17, p. 9483–9494, 2024. Citado 4 vezes nas páginas 2, 7, 10 e 15.
- GHAZALI, M.; TAHAR, A. B. Cost optimization of the GI/GI/n queueing model. **International Journal of Applied Mathematics**, v. 55, n. 5, 2025. Citado na página 9.
- HAVIV, M.; RAVNER, L. A survey of queueing systems with strategic timing of arrivals. **Queueing Systems**, Springer, v. 99, n. 1, p. 163–198, 2021. Citado na página 5.
- JUAN, A. A. *et al.* A review of the role of heuristics in stochastic optimisation: From metaheuristics to learnheuristics. **Annals of Operations Research**, Springer, v. 320, n. 2, p. 831–861, 2023. Citado na página 1.

KENDALL, D. G. Stochastic processes occurring in the theory of queues and their analysis by the method of embedded Markov chains. **Annals Mathematical Statistics**, v. 24, p. 338–354, 1953. Citado 3 vezes nas páginas 2, 5 e 6.

MARTÍ, R.; REINELT, G. Exact and heuristic methods in combinatorial optimization. **Applied Mathematical Sciences**, Springer, 2022. Citado na página 1.

OLIVEIRA, M. C. *et al.* Optimal buffer and server allocation in markovian queueing networks by particle swarm algorithm. **Sigmae**, v. 14, n. 5, p. 01–17, 2025. Citado na página 7.

RARDIN, R. L. **Optimization in operations research**. [S.l.]: Prentice Hall Upper Saddle River, NJ, 1998. v. 166. Citado na página 4.

RINGUEST, J. L. **Multiobjective optimization: behavioral and computational considerations**. [S.l.]: Springer Science & Business Media, 2012. Citado na página 3.

SMITH, J. M. $M/G/c/K$ blocking probability models and system performance. **Performance Evaluation**, v. 52, n. 4, p. 237–267, 2003. Citado na página 13.

SMITH, J. M. Optimal design and performance modelling of $M/G/1/K$ queueing systems. **Mathematical and Computer Modelling**, v. 39, n. 9-10, p. 1049–1081, 2004. Citado na página 13.

SMITH, J. M.; CRUZ, F. R. B. The buffer allocation problem for general finite buffer queueing networks. **IIE Transactions**, v. 37, n. 4, p. 343–365, 2005. Citado na página 7.

SOUZA, G. L. **Uma nova formulação para otimização multi-objetivo em redes de filas finitas gerais e com único servidor**. Tese (Mestrado em Ciência da Computação) — Universidade Federal de Ouro Preto, 2020. Citado na página 6.

STOIBER, N.; KROMOSER, B. Topology optimization in concrete construction: a systematic review on numerical and experimental investigations. **Structural and Multidisciplinary Optimization**, Springer, v. 64, n. 4, p. 1725–1749, 2021. Citado na página 4.

YANG, X. S. **Engineering Optimization: An Introduction with Metaheuristic Applications**. 1st. ed. [S.l.]: Wiley Publishing, 2010. ISBN 0470582464, 9780470582466. Citado 2 vezes nas páginas 1 e 3.