



UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA

DEST
DEPARTAMENTO DE
ESTATÍSTICA

José Victor Sá Santos

Otimização Multiobjetivo de Redes de Filas Finitas de Servidor Único por Meio de um Algoritmo de Enxame de Partículas

Belo Horizonte

2023

José Victor Sá Santos

Otimização Multiobjetivo de Redes de Filas Finitas de Servidor Único por Meio de um Algoritmo de Enxame de Partículas

Monografia apresentada ao Curso de Especialização em Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Especialista em Estatística

Orientador: Prof. Dr. Frederico Rodrigues Borges da Cruz

Coorientador: Anderson Ribeiro Duarte

Belo Horizonte

2023

2023, José Victor Sá Santos.
Todos os direitos reservados

Santos, José Victor Sá.

S237a Otimização multiobjetivo de redes de filas finitas de servidor único por meio de um algoritmo de enxame de partículas / José Victor Sá Santos — 2023.
22.f. il.

Orientador: Frederico Rodrigues Borges da Cruz.
Coorientador: Anderson Ribeiro Duarte
Monografia (especialização) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: 20-22.

1. Estatística. 2. Teoria das filas. 3. Otimização multiobjetivo. I. Cruz, Frederico Rodrigues Borges da. II. Duarte, Anderson Ribeiro. III. Universidade Federal de Minas Gerais Instituto de Ciências Exatas, Departamento de Estatística. IV. Título.

CDU 519.2 (043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa CRB 6/1510 Universidade Federal de Minas Gerais – ICEX



Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

ATA DO 308ª. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE JOSÉ VÍCTOR SÁ SANTOS.

Aos quinze dias do mês de setembro de 2023, às 14:00 horas, com utilização de recursos de videoconferência à distância, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora, homologada pela Comissão do Curso de Especialização em Estatística, para julgar a apresentação do trabalho de fim de curso do aluno **José Victor Sá Santos**, intitulado: “**Otimização Multiobjetivo de Redes de Filas Finitas de Servidor Único por Meio de um Algoritmo de Enxame de Partículas**”, como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, o Presidente da Comissão, Professor Frederico Rodrigues Borges da Cruz (orientador), após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra ao candidato para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores, com a respectiva defesa do candidato. Após a defesa, os membros da banca examinadora reuniram-se sem a presença do candidato e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: o candidato foi considerado **aprovado**, por unanimidade, condicional às modificações sugeridas pela Comissão Examinadora, no prazo de 30 (trinta) dias, a partir da data de hoje. O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da Comissão Examinadora. Belo Horizonte, 15 de setembro de 2023.

Documento assinado digitalmente
 FREDERICO RODRIGUES BORGES DA CRUZ
Data: 15/09/2023 15:04:44 -0300
Verifique em <https://validar.iti.gov.br>

Prof. Frederico Rodrigues Borges da Cruz (Orientador)
Departamento de Estatística / UFMG

Documento assinado digitalmente
 ANDERSON RIBEIRO DUARTE
Data: 15/09/2023 14:43:41 -0300
Verifique em <https://validar.iti.gov.br>

Prof. Anderson Ribeiro Duarte (Coorientador)
Departamento de Estatística / UFOP

Documento assinado digitalmente
 HELGEM DE SOUZA RIBEIRO MARTINS
Data: 15/09/2023 14:48:17 -0300
Verifique em <https://validar.iti.gov.br>

Prof. Helgem de Souza Ribeiro Martins
Departamento de Estatística / UFOP

RESUMO

As longas e demoradas filas são uma experiência comum e ocorrem em diversas situações do cotidiano. Modelos matemáticos de filas são frequentemente utilizados para entender e melhorar esses sistemas. Este estudo foca na otimização de redes de filas finitas, com interesse em maximizar a taxa de atendimento (o *throughput*), minimizar as capacidades e abordar o *trade-off* entre a alocação de capacidades e o *throughput*. É descrito um método heurístico, baseado no algoritmo de enxame de partículas, para encontrar um conjunto Pareto de soluções sub-ótimas. A abordagem é ilustrada com redes de filas $M/G/1/k$, na notação de Kendall, via utilização de uma ferramenta clássica para análise de desempenho, o método de expansão generalizado (GEM). Experimentos mostram a eficácia e a eficiência do algoritmo, com alta qualidade de soluções e baixos tempo de processamento. Investigações futuras podem explorar diferentes configurações de redes e aplicações em cenários reais. O estudo contribui para a otimização de sistemas de filas e suas aplicações práticas.

Palavras-chave: redes de filas; otimização multiobjetivo; medidas desempenho; enxame de partículas.

ABSTRACT

Long and time-consuming queues are a common experience and occur in various everyday situations. Mathematical queue models are often applied to understand and to improve these systems. This study focuses on optimizing finite queueing network systems, aiming to maximize throughput, minimize the overall capacity, and to address the trade-off between capacity allocation and throughput. A heuristic method based on the particle swarm algorithm is presented to identify a Pareto set of sub-optimal solutions. The approach is illustrated with networks of $M/G/1/k$ queues, in Kendall notation, utilizing a classical performance analysis tool, the generalized expansion method (GEM). Experiments demonstrate efficacy and computational efficiency in terms of high quality solutions and low CPU times. Future investigations could explore diverse network configurations and real-world applications. The study contributes to the optimization of queueing systems and their practical implementations.

Keywords: queueing networks; multiobjective optimization; performance measures; particle swarm.

LISTA DE FIGURAS

1	Uma rede de filas em topologia mista, adaptada de Smith e Cruz (2005)	11
2	Pontos dominados (■) e não dominados (●)	16
3	Ilustração da distância de aglomeração	16
4	Esquema ilustrativo de utilização do método de expansão generalizado .	19
5	Topologias testadas	23
6	Evolução das fronteiras de Pareto em função no número de ciclos	25

LISTA DE TABELAS

1	Algumas soluções dominantes em função da topologia e do s^2	26
---	---	----

SUMÁRIO

1	Introdução	9
2	Aspectos Metodológicos	13
2.1	Formulação com Objetivo Único	13
2.2	Formulação Multiobjetivo	14
2.3	Abordagem de Otimização Multiobjetivo	14
2.4	Avaliação de Desempenho	17
2.5	Algoritmo de Enxame de Partículas	20
3	Resultados e Discussão	23
4	Conclusões e Observações Finais	27
	REFERÊNCIAS	28

1 INTRODUÇÃO

As longas e demoradas filas têm sido experimentadas por todos, por ocorrer em diversas situações cotidianas, como no trânsito urbano engarrafado, em caixas de estabelecimentos de varejo, agências bancárias, serviço público, dentre inúmeras outras situações semelhantes. Tal fenômeno, que pode ser representado matematicamente, de forma aproximada, pelos modelos de filas, ocorre porque a demanda por atendimento se torna maior do que a capacidade do sistema de serviços atender a tal demanda. Uma solução simplista seria aumentar a capacidade de serviço ao máximo possível, mas as restrições de orçamento e espaço tornam tal escolha inviável.

Uma forma instrutiva para descrever um sistema de filas é através de um modelo de chegadas de clientes que esperam pelo serviço desejado e posteriormente abandonam o sistema após terem suas demandas atendidas. Trata-se de um cenário em que há incertezas quanto ao fluxo de usuários e quanto ao seu atendimento e diversos exemplos práticos servem para reconhecer imediatamente tal situação, com a inclusão de uma generalização natural, que é composição de filas em *redes*, em que a i -ésima fila tem a taxa de chegada λ_i e a taxa de atendimento μ_i . De fato, é possível encontrar exemplos relevantes de redes de filas em artigos recentes na área da estatística ([CRUZ; DUARTE; SOUZA, 2018](#); [MARTINS et al., 2019](#); [SOUZA et al., 2023](#)).

Situações de filas com capacidade limitada, para um determinado serviço, resultam no que nominamos por *filas finitas*. No caso de filas finitas com espaço total para k clientes, P_k denota a probabilidade de encontrar k usuários no sistema, com a inclusão daqueles que já estão em atendimento. Diante disso, P_k é a probabilidade de bloqueio da fila em estudo. Quando um cliente chega em busca do serviço e todos os servidores e as posições de espera estão ocupados, o cliente é bloqueado pelo sistema. Por razões óbvias, altas probabilidades de bloqueio implicam na ineficiência do sistema de filas ([CRUZ; DUARTE; WOENSEL, 2008](#)).

A novidade deste estudo consiste principalmente no exame do problema de otimização de redes de filas finitas, por meio de um método heurístico eficaz para maximizar a taxa de atendimento (o *throughput*) e simultaneamente minimizar as capacidades, em redes acíclicas compostas por m filas $M/G/1/k$, de acordo com a clássica notação de [Kendall \(1953\)](#), ou seja, redes de filas com chegadas markovianas (M), tempos de atendimento geral (G), servidor único (1) e capacidade máxima de k clientes simultâneos, com a inclusão daquele em serviço. Para obter o *throughput* máximo Θ e a capacidade total mínima $\sum_i k_i$, em uma dada topologia e sob uma taxa de chegada conhecida λ_i , o procedimento para o processo de otimização busca as coordenadas ótimas do vetor $\mathbf{k} = (k_1, k_2, \dots, k_m)$, que determinem a configuração ideal para a rede de filas.

A otimização em redes de filas finitas diz respeito a diversos aspectos da vida real, com a possibilidade de ajudar a compreender e a melhorar variados sistemas presentes no cotidiano das pessoas. Entre eles, podem ser citados exemplos de aplicações na indústria de alimentos ([MARTINS et al., 2019](#)), em sistemas de saúde ([OSORIO; BIERLAIRE, 2009](#)), no tráfego urbano ([SMITH; CRUZ; WOENSEL, 2010](#)), em sistemas baseados na *web* ([MENASCÉ, 2002](#)) e vários outros.

Um exemplo específico do tipo de rede de filas $M/G/1/k$ enfocado neste trabalho pode ser visualizado na Figura 1. Trata-se de uma topologia complexa de rede de filas que envolve filas em séries, divisões do fluxo para mais filas e fusões entre filas. A entrada no sistema é única, através da primeira fila do sistema, e continua pelo sistema. Para situações de divisão, as probabilidades de roteamento são supostamente conhecidas.

Existe um *trade-off* crítico entre a alocação de capacidades em cada fila do sistema e o *throughput* do sistema de filas. É razoável admitir que, em geral, quanto maior a alocação de capacidades, maior será o *throughput*. Por outro lado, a alocação de capacidades é bastante onerosa. O objetivo maior é então minimizar a alocação de tais recursos, mas ainda assim conseguir obter um *throughput* satisfatório para o

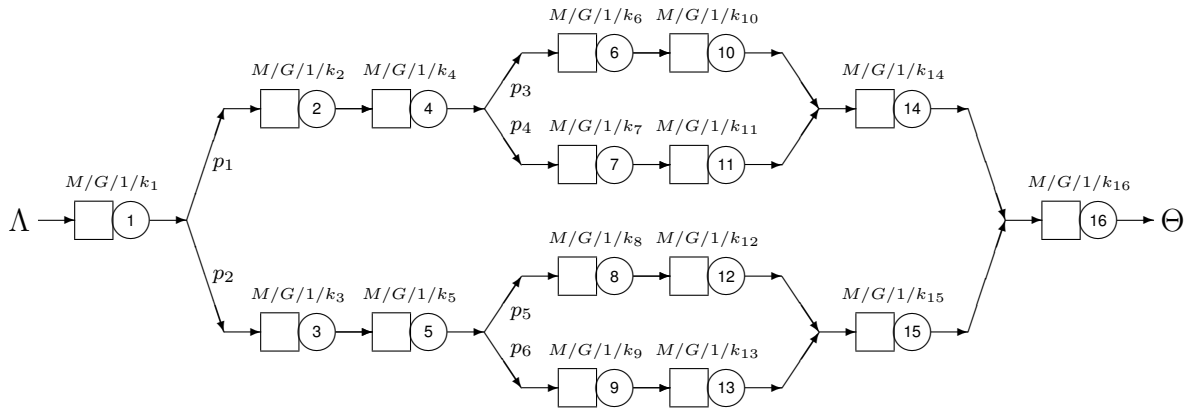


Figura 1: Uma rede de filas em topologia mista, adaptada de [Smith e Cruz \(2005\)](#)

sistema.

Uma abordagem heurística é proposta para a busca de um conjunto Pareto de soluções sub-ótimas. O método produz um conjunto de soluções eficientes para mais de um objetivo presente na função objetivo. Com a abordagem proposta, o decisor pode avaliar o efeito de cada escolha. A abordagem multiobjetivo também permite que o usuário melhore um objetivo (por exemplo, aumente o *throughput*) enquanto sacrifica (piora) outro objetivo (por exemplo, aumenta as capacidades alocadas, o que acarreta um aumento nos custos). Um algoritmo de otimização de enxame de partículas multiobjetivo foi utilizado. Os algoritmos heurísticos inspirados na natureza são particularmente adequados para problemas difíceis com apenas um objetivo, mas sua versão multiobjetivo também se mostrou eficaz em uma variedade de configurações com funções objetivos e restrições particularmente complicadas, com ótimo desempenho computacional ([LEONG; YEN, 2008](#); [HAJIPOUR; PASANDIDEH, 2012](#); [SHARAFI; ELMEKKAWY, 2014](#); [DENG et al., 2017](#); [AZIMI; ASADOLLAHI, 2019](#); [OLIVEIRA et al., 2019](#)).

O restante do texto está assim organizado. O Capítulo 2 discute formulações de programação matemática para a otimização das redes de filas e apresenta o algoritmo de enxame de partículas para a otimização multiobjetivo. No Capítulo 3, os resultados de um conjunto abrangente de experimentos são apresentados e demonstram a eficácia e eficiência da abordagem. Por fim, o Capítulo 4 conclui o texto, com

considerações finais e proposições para trabalhos futuros na área.

2 ASPECTOS METODOLÓGICOS

Este estudo apresenta uma investigação a respeito de problemas de otimização aplicados à busca da relação eficiente entre consumo de recursos e alto desempenho. Diante disso, é relevantes a discussão sobre formulações matemáticas adequadas ao problema e sobre ferramentas algorítmicas úteis para a resolução destes problemas. Estes aspectos serão melhor elucidados a seguir, primeiro com a formulação do problema e em seguida pela apresentação dos algoritmos envolvidos.

2.1 FORMULAÇÃO COM OBJETIVO ÚNICO

O propósito aqui é a apresentação de algoritmos para otimizar (minimizar) a alocação total de capacidade (ou seja, $\sum_i k_i$) de uma rede acíclica de filas $M/G/1/k$, enquanto simultaneamente otimiza (maximiza) a taxa de atendimento (ou seja, o *throughput*, Θ). O algoritmo é muito dependente da formulação de programação matemática. Assim, inicialmente é apresentada a formulação mais comum para o problema de alocação de capacidades (ou BAP, do inglês *buffer allocation problem*).

O BAP é definido em um dígrafo $\mathcal{D}(V, A)$, em que V é um conjunto finito de m vértices (filas), e A é um conjunto finito de arestas (conexões entre as filas). O BAP, em sua formulação primal (SMITH; CRUZ, 2005), é apresentado a seguir:

$$\text{minimize } \sum_{i=1}^m c_i k_i, \quad (1)$$

sujeito a:

$$\begin{aligned} \Theta(\mathbf{k}) &\geq \Theta_{\min}, \\ k_i &\in \mathbb{N}, \forall i \in \{1, 2, \dots, m\}, \end{aligned} \quad (2)$$

que minimiza os custos totais de alocação de capacidades $\sum_{i=1}^m c_i k_i$, sujeito a uma taxa de atendimento $\Theta(\mathbf{k})$ superior a um limiar dado $\Theta_{\min}$ e a capacidades k_i inteiras e positivas.

Esta formulação é a mais usual na literatura, em que o *throughput* é mode-

lado como uma restrição. Uma desvantagem dessa abordagem é que se a restrição de *throughput* for relaxada, a determinação de um limiar $\Theta_{\min}$ adequado não é uma tarefa trivial. Assim, embora a formulação BAP possibilite o desenvolvimento de algoritmos eficientes para sua resolução, este estudo considera um algoritmo baseado na formulação multiobjetivo apresentada a seguir.

2.2 FORMULAÇÃO MULTIOBJETIVO

O problema de otimização de redes $M/G/1/k$ apresentado na seção anterior pode ser descrito por meio de uma formulação de programação matemática multiobjetivo, que compreende a minimização da alocação total de capacidade, simultaneamente com a maximização do *throughput*. Este problema da rede de filas multiobjetivo pode ser formulado com se segue:

$$\text{minimize } F(\mathbf{k}, \boldsymbol{\mu}) = \left[f_1(\mathbf{k}), f_2(\mathbf{k}) \right], \quad (3)$$

sujeito a:

$$k_i \in \mathbb{N}, \forall i \in \{1, 2, \dots, m\}, \quad (4)$$

em que $f_1(\mathbf{k}) = \sum_{i=1}^m k_i$ representa a alocação total de capacidade e $f_2(\mathbf{k}) = -\Theta(\mathbf{k})$ representa o negativo da taxa de atendimento (*throughput*). Observe que o sinal negativo no *throughput* possibilita a formulação do problema em termos apenas de minimização, sem perda de qualquer generalidade.

2.3 ABORDAGEM DE OTIMIZAÇÃO MULTIOBJETIVO

A otimização multiobjetivo aborda o problema de busca de soluções ótimas em que várias funções objetivo concorrentes estão em interação. Para efeito de definição, os contextos de maximização e de minimização são análogos. Um problema de otimização multiobjetivo compreende um par de objetos $\mathcal{X}$ e F , definidos como:

$$\begin{aligned} &\text{minimize } F(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_\ell(\mathbf{x})], \\ &\text{sujeito a: } \mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathcal{X}, \end{aligned} \quad (5)$$

em que $\mathbf{x} \in \mathcal{X}$ é a *variável de decisão*, $\mathcal{X}$ é o *conjunto de soluções viáveis* e $\mathcal{F} = F(\mathcal{X})$ é o *espaço de objetivos*, com ℓ funções objetivo. O interesse na otimização multiobjetivo é encontrar as soluções que “minimizem” $F(\mathbf{x})$ no contexto de otimalidade de Pareto (MOREIRA; PAQUETE, 2019).

Dado $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ e a relação $\prec$, definida como

$$F(\mathbf{x}) \prec F(\mathbf{x}') \iff F(\mathbf{x}) \leq F(\mathbf{x}') \text{ e } F(\mathbf{x}) \neq F(\mathbf{x}'), \quad (6)$$

em que $F(\mathbf{x}) \leq F(\mathbf{x}')$ se e somente se $f_i(\mathbf{x}) \leq f_i(\mathbf{x}')$, para todos os valores $i \in \{1, \dots, \ell\}$ e $F(\mathbf{x}) \neq F(\mathbf{x}')$ se e somente se $\exists i \in \{1, \dots, \ell\}$, tal que $f_i(\mathbf{x}) \neq f_i(\mathbf{x}')$. Uma solução factível $\mathbf{x}^* \in \mathcal{X}$ é uma *solução Pareto-ótimo* do problema de otimização multiobjetivo dado pela Eq. (5) se não houver $\mathbf{x} \in \mathcal{X}$ tal que $F(\mathbf{x}) \prec F(\mathbf{x}^*)$. O conjunto de todas as soluções Pareto-ótimas é chamado de *conjunto ótimo de Pareto*, e o subconjunto da imagem pela função F , definido pelas soluções Pareto-ótimas, é chamado de *fronteira de Pareto*. Diz-se que a solução $\mathbf{x}$ domina a solução $\mathbf{x}'$ e/ou que $F(\mathbf{x})$ domina $F(\mathbf{x}')$ se $F(\mathbf{x}) \prec F(\mathbf{x}')$.

Por exemplo, a Figura 2 exibe os pontos para um problema de minimização de duas dimensões, em que o ponto V é dominado pelo ponto I, mas não pelos pontos II, III e IV. Além disso, o ponto VI é dominado pelos pontos I, II e III, mas não pelo ponto IV. A melhor fronteira inclui os pontos I a IV e é uma aproximação para o conjunto de Pareto, que é o conjunto de pontos que não são dominados por outros pontos.

Para realizar o *elitismo*, que é a seleção dos (melhores) pontos que passarão para a próxima iteração, o *algoritmo de ordenação rápida não-dominada* (do inglês, *fast non-dominated sorting algorithm*) foi empregado. Maiores detalhes podem ser encontrados em Deb et al. (2002), mas essencialmente este algoritmo separa os in-

divíduos da população em várias camadas ou fronteiras $\mathcal{F}_i$, de modo que as soluções em $\mathcal{F}_1$ sejam não dominadas e todas as soluções em uma dada fronteira $\mathcal{F}_i$, $i > 1$ sejam dominadas por pelo menos uma solução em $\mathcal{F}_{i-1}$ e não sejam dominadas por nenhuma solução na fronteira $\mathcal{F}_j$, $j \geq i$.

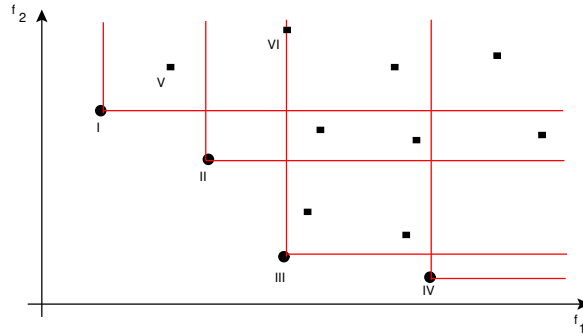


Figura 2: Pontos dominados (■) e não dominados (●)

A seleção é realizada escolhendo-se em sequência os pontos de cada fronteira não dominada ($\mathcal{F}_1, \mathcal{F}_2, \dots$) até que seja obtido o número de indivíduos necessários para a próxima iteração. Algum critério precisa ser aplicado se, após a adição de um grupo de indivíduos de $\mathcal{F}_i$, o número máximo de indivíduos for excedido. O algoritmo calcula uma medida de diversidade, denominada *distância de aglomeração* (do inglês, *crowding distance*), conforme definido por Deb et al. (2002), que garante a maior diversidade possível. Assim, apenas os pontos com a maior distância de aglomeração são mantidos para futuras iterações, conforme mostrado na Figura 3.

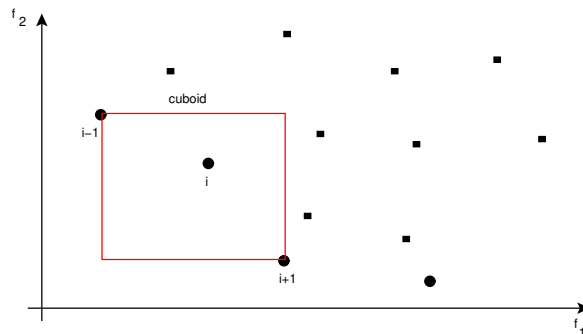


Figura 3: Ilustração da distância de aglomeração

Ao considerar a formulação matemática especificada para este problema, a sequência do procedimento exige calcular, para cada solução sob estudo, o valor das

funções objetivo descritas nas Eq. (3) e (4). Particularmente, o cálculo de $f_2(\mathbf{x})$ não é trivial e demanda a aplicação de mecanismos específicos de avaliação de desempenho. É necessária uma técnica eficaz de aproximação das probabilidades de bloqueio nas filas da rede, ou seja, a probabilidade de um usuário chegar e todos os lugares de espera na fila estarem ocupados. Além disso, um método para obtenção do *throughput* da rede de filas, denominado *método de expansão generalizado* (ou GEM, do inglês, *generalized expansion method*) e desenvolvido por [Kerbach e Smith \(1987\)](#), é utilizado.

2.4 AVALIAÇÃO DE DESEMPENHO

Em uma única fila $M/G/1/k$, a estimativa da probabilidade de bloqueio P_k pode ser obtida através de uma forma fechada, computacionalmente eficiente e precisa. O método, proposto por [Smith \(2004\)](#), é baseado em uma aproximação de dois momentos de [Kimura \(1996\)](#):

$$P_k = \frac{\rho \left(\frac{2 + \sqrt{\rho s^2 - \sqrt{\rho} + 2(k-1)}}{2 + \sqrt{\rho s^2 - \sqrt{\rho}}} \right) (\rho - 1)}{\rho \left(2 \frac{2 + \sqrt{\rho s^2 - \sqrt{\rho} + (k-1)}}{2 + \sqrt{\rho s^2 - \sqrt{\rho}}} \right) - 1}, \quad (7)$$

em que $\rho < 1$ é uma condição que deve sempre estar atendida, em que ρ é a intensidade de tráfego, ou seja, a taxa de utilização do sistema, definida como a razão entre a taxa efetiva de chegada na fila e a sua taxa de serviço, $\rho = \lambda/\mu$. Além disso, $s^2 = \mathbb{V}(T_s)/\mathbb{E}^2(T_s)$ é o quadrado do coeficiente de variação do tempo de serviço (T_s). Vários estudos anteriores confirmam que a aproximação de P_k dada pela Eq. (7) é precisa para uma ampla faixa de valores de k ([SMITH; CRUZ, 2005](#); [CRUZ; DUARTE; WOENSEL, 2008](#)).

Para filas únicas, uma fração P_k das chegadas não entra no sistema, ou seja, é bloqueada na entrada. Assim, P_k representa a probabilidade de um cliente chegar quando não houver mais capacidade disponível. Portanto, apenas a fração $(1 - P_k)$ das chegadas será atendida pela fila ([GROSS et al., 2009](#)). Isto é, na prática, apenas

a fração $(1 - P_k)$ dos clientes não encontra o sistema bloqueado. Também, a taxa de chegada é aproximadamente markoviana, ou seja, os tempos entre chegadas seguem aproximadamente uma distribuição exponencial com taxa λ . Essa informação é determinante para o cálculo do *throughput* dessa fila única (SMITH; CRUZ, 2005). Consequentemente, essa situação resulta em um *throughput* igual a $\Theta = \lambda(1 - P_k)$.

Os problemas de otimização em redes de filas tem sido abordados por diversas perspectivas. Há abordagens por meio de métodos de otimização direta sem restrições, como o algoritmo de Powell (MARTINS et al., 2019), por algoritmos genéticos (CRUZ et al., 2012), e o recozimento simulado (do inglês *simulated annealing*) (CRUZ; DUARTE; SOUZA, 2018). Essas abordagens maximizam o *throughput* (Θ), determinado aproximadamente pelo método de expansão generalizado (GEM), que é uma ferramenta apropriada para determinação do *throughput* em filas configuradas em *redes*.

O GEM é um algoritmo usado com sucesso para estimar o desempenho de redes acíclicas de fila finita configuradas arbitrariamente em diversos estudos anteriores (CRUZ; DUARTE; WOENSEL, 2008; CRUZ et al., 2012; CRUZ; DUARTE; SOUZA, 2018). No processo de estimação, o GEM atualiza as medidas de desempenho do sistema em tentativas repetidas. O método considera o efeito de atraso gerado por diversos possíveis bloqueios ocorridos no fluxo de clientes ao longo da rede de filas. O GEM resolve um conjunto de equações não lineares simultâneas por meio de procedimentos iterativos. Isso leva a uma melhoria considerável na precisão da estimativa das medidas de desempenho da rede de filas. O método é uma combinação de decomposição *nó a nó* e tentativas repetidas, nas quais cada fila é analisada separadamente e as correções são feitas para levar em conta os efeitos entre as filas da rede.

Conforme descrito em detalhes na literatura (KERBACHE; SMITH, 1987), o GEM cria, para cada nó finito j , um vértice auxiliar (h_j) que é modelado como uma fila $M/G/\infty$, conforme mostrado na Figura 4. Para cada entidade no sistema, o vértice

j pode ser bloqueado (com probabilidade P_{k_j}) ou pode ser desbloqueado (com probabilidade $1 - P_{k_j}$). Quando ocorre o bloqueio, as entidades são redirecionadas para o vértice h_j e são atrasadas enquanto o nó j estiver ocupado. O vértice h_j registra o tempo que uma entidade tem que esperar, com uma taxa de serviço μ'_h , dada pelo GEM, antes de entrar no vértice j , e atualiza de acordo com a taxa de chegada efetiva proveniente do vértice i para o vértice j , $\lambda_{\text{eff}} = \lambda_i(1 - P_{k_i})$.

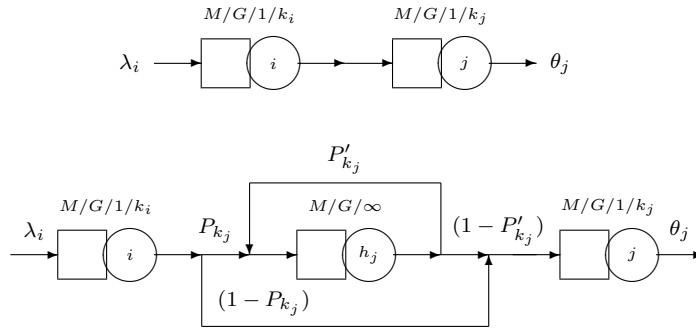


Figura 4: Esquema ilustrativo de utilização do método de expansão generalizado

O objetivo final do GEM é fornecer atualizações das taxas de serviço dos nós da seguinte forma:

$$\tilde{\mu}_i^{-1} = \mu_i^{-1} + P_{k_j}(\mu'_h)^{-1}, \quad (8)$$

de tal forma que a partir da Eq. (8) as taxas de serviço μ_i podem ser atualizadas para todos os nós ao longo de ρ e consequentemente P_k , da Eq. (7).

O GEM calcula iterativamente as atualizações das medidas de desempenho do sistema. O método leva em consideração o efeito de atraso gerado por diversos possíveis bloqueios ocorridos no fluxo de clientes ao longo das filas. É importante notar que o cálculo de P_k , por meio da aproximação proposta na Eq. (7), é dependente do conhecimento da taxa de chegada da fila, λ . Para a fila frontal de uma rede (veja Figura 1), a taxa de chegada, λ , é conhecida. No entanto, quais são as taxas de chegada para as filas subsequentes? O procedimento aplicado para obter essas taxas de chegada em muitos estudos também leva em conta as aproximações produzidas com o uso do GEM.

Observe que P_{k_i} é calculado pelo GEM (SMITH, 2004) e depende de λ_i , μ_i e k_i . O valor k_i é variável de decisão do problema de otimização, mas a taxa de chegada λ_i depende da taxa de *throughput* da fila anterior. Sem perda de generalidade, em uma rede de filas tandem, os cálculos realizados neste estudo sempre assumem que a taxa de chegada na i -ésima fila é dependente da $(i - 1)$ -ésima fila anterior, dada por:

$$\lambda_i = \lambda_{i-1}(1 - P_{K_{i-1}}), \quad (9)$$

em que $i \in \{2, \dots, m\}$ e λ_1 é considerado a chegada da rede de filas externa, ou seja, $\lambda_1 = \Lambda$.

2.5 ALGORITMO DE ENXAME DE PARTÍCULAS

Um algoritmo de otimização de enxame de partículas multiobjetivo (MOPSO, do inglês *multiobjective particle swarm optimization*) é proposto para resolver a formulação multiobjetivo do problema de otimização de redes de filas aqui estudado, descrito pelas Eqs. (3) e (4). O MOPSO proposto estende o algoritmo PSO de objetivo único proposto por Kennedy e Eberhart (1995). Cada partícula deve representar uma possível solução para a alocação de recursos (capacidades k_i) que otimize a rede de filas finitas em estudo. Assim, nesta formulação particular, cada partícula pode ser representada pelas variáveis $(x_1, \dots, x_m) = (k_1, k_2, \dots, k_m)$.

É importante destacar aqui que o problema de otimização multiobjetivo abordado é um problema com variáveis de decisão inteiras. Assim, uma estratégia de adaptação de partículas deve ser definida. De fato, as alterações nas capacidades são realizadas e, em seguida, os valores inteiros positivos são usados, pois $k_i \geq 1$ é sempre respeitado. Da mesma forma, as restrições associadas às taxas de serviço também são respeitadas, pois é necessário garantir que $\rho < 1$. De outra forma, a taxa de chegada da fila λ_i deve ser estritamente menor que a taxa de serviço μ_i . Essas considerações garantem a viabilidade das soluções investigadas.

Defina s o tamanho do enxame (população de partículas) utilizado. Cada partícula i , com $1 \leq i \leq s$, tem os seguintes atributos:

- posição, $x_i = (x_{i,1}, x_{i,2}, \dots, x_{i,m})$;
- velocidade, $v_i = (v_{i,1}, v_{i,2}, \dots, v_{i,m})$;
- melhor posição pessoal, p_i ;
- melhor posição global, g_i .

A abordagem MOPSO proposta para otimização da rede de filas pode ser descrita pela execução da seguinte sequência de etapas:

Passo 1: Inicializar:

- população de partículas, x_i ;
- velocidade das partículas, v_i ;
- melhor posição pessoal, $p_i = x_i$;
- melhor posição global entre partículas, $g_i = x_i$;
- contador de iterações;

Passo 2: Armazenar as partículas não dominadas de x_i no arquivo A ;

Passo 3: Calcular a *distância de aglomeração* das partículas armazenadas em A e ordená-las de forma decrescente;

Passo 4: Selecionar aleatoriamente uma das partículas não dominadas de A e armazenar a posição em g_i ;

Passo 5: Atualizar a velocidade e a posição das partículas de acordo com as Eqs. (10) e (11), respectivamente:

$$v_i^{t+1} = w^t + r_1(p_i - x_i^t) + r_2(g_i - x_i^t), \quad (10)$$

$$x_i^{t+1} = x_i^t + v_i^{t+1}; \quad (11)$$

Como as variáveis de decisão (as capacidades alocadas) são inteiras, a posição deve ser atualizada de acordo com a Eq. (12):

$$x_i^{t+1} = \text{int}(x_i^t + v_i^{t+1}); \quad (12)$$

Passo 6: Atualizar a melhor posição de cada partícula, p_i ; se a posição corrente dominar a posição p_i na memória, então alterar p_i para a posição corrente.

Passo 7: Retorne para o Passo 2 até que o critério de parada seja satisfeito.

Os parâmetros de funcionamento do MOPSO e seus valores foram definidos da seguinte forma: r_1 e r_2 são números aleatórios positivos com distribuição uniforme pertencentes ao intervalo $[0; 1,0]$, $w = 0,4$ é o peso de inércia.

Maiores detalhes sobre a implementação dos algoritmos MOPSO podem ser encontrados na literatura. O MOPSO que foi descrito é uma adaptação da implementação clássica proposta por [Coello-Coello e Lechuga \(2002\)](#). No entanto, versões simplificadas podem ser encontradas, como a apresentada por [Trivedi, Varshney e Ramteke \(2019\)](#) e versões mais sofisticadas e melhoradas, como as apresentadas por [Fan et al. \(2017\)](#), [Jia e Zhu \(2017\)](#). Versões com formulações de programação matemática de inteiros mistos também estão disponíveis, como a apresentada por [Zhao et al. \(2013\)](#).

3 RESULTADOS E DISCUSSÃO

Os algoritmos de otimização foram implementados em FORTRAN. O código aqui utilizado foi gentilmente cedido por [Cruz, Duarte e Souza \(2018\)](#) e [Souza et al. \(2020\)](#), e está disponível a pedido diretamente com tais autores, para fins educacionais e de pesquisa. Os experimentos computacionais foram executados em um computador pessoal, processador Intel(R) Core(TM) i3-2310M 2,10 GHz rodando Windows 10 Pro 64 bits, com 6,00GB de RAM.

O algoritmo de enxame de partículas foi testado nas redes de filas configuradas nas topologias elementares série, divisão e fusão, conforme apresentado na Figura 5. As chegadas são exponenciais (markovianas) com taxa $\Lambda = 5,0$. Três diferentes valores para o quadrado do coeficientes de variação do tempo de serviço geral foram analisados, $s^2 = \{0.5, 1.5\}$, representando serviços hipoexponenciais e hiperexponenciais, respectivamente, mas a qualidade das soluções não se alterou significativamente. O número de partículas no enxame foi definido como 400, e o número máximo de interações do algoritmo foi definido como 4.000, com tempo de execução em torno de 2 minutos, para os casos testados.

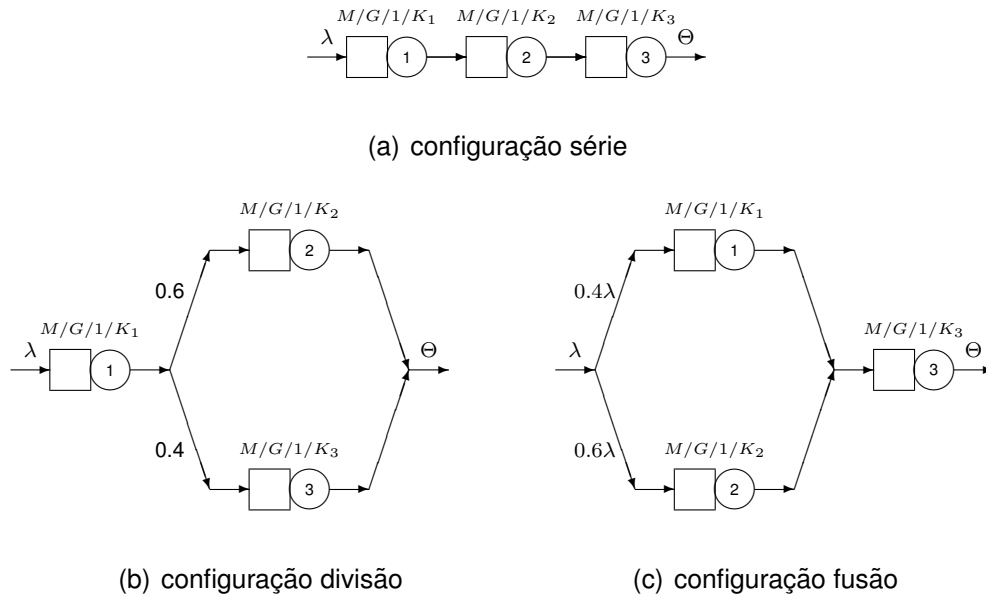


Figura 5: Topologias testadas

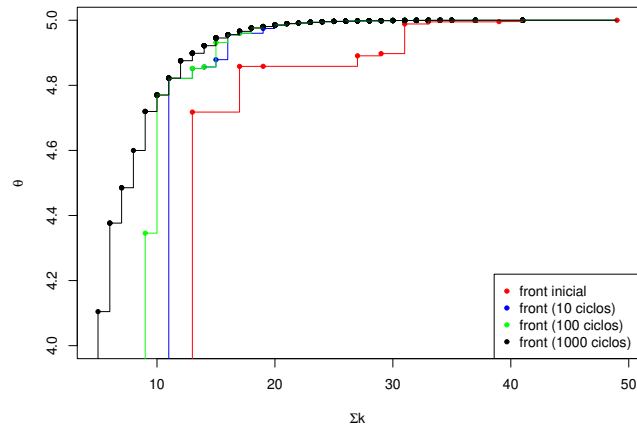
A Figura 6 apresenta a evolução das fronteiras de Pareto, em função do

número de ciclos, para as três topologias testadas, o que demonstra uma capacidade de geração de soluções satisfatórias, independentemente da topologia, o que é bastante promissor.

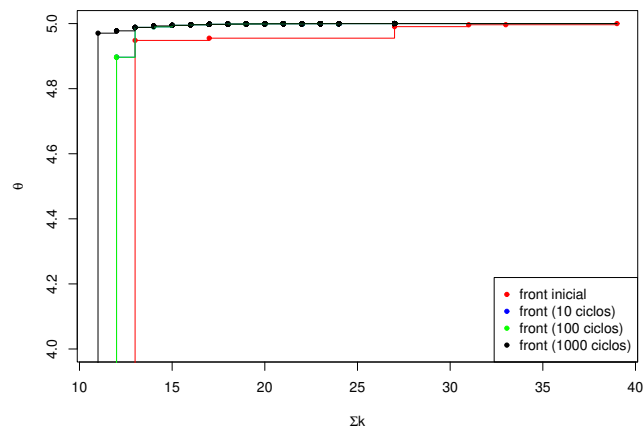
Também de interesse é observar os resultados no espaço das soluções, conforme apresentado na Tabela 1. Nota-se aqui a multiplicidade de soluções *eficientes* (isto é, soluções não dominadas) que a abordagem multiobjetivo produz. De fato, nessas soluções não é possível produzir o mesmo *throughput* com menor uso de capacidade total alocada. Também não é possível manter a capacidade total alocada e ampliar o *throughput*. A vantagem é que o analista tem como escolher a solução eficiente de interesse, após uma análise de *trade-off* entre *throughput* e capacidade total alocada.

Analisando-se as soluções apresentadas, nota-se que com o aumento do quadrado do coeficiente de variação do tempo de serviço, s^2 , há um aumento na alocação de capacidades, para uma dada taxa de saída, o que é esperado, para dar conta da maior variabilidade da taxa de atendimento. Além disso, para as filas em série, nota-se uma alocação mais homogênea, mas com tendência de alocação maior nas filas das extremidades (fila de entrada e fila de saída), fenômeno conhecido com *efeito bacia* (do inglês, *bowl effect*).

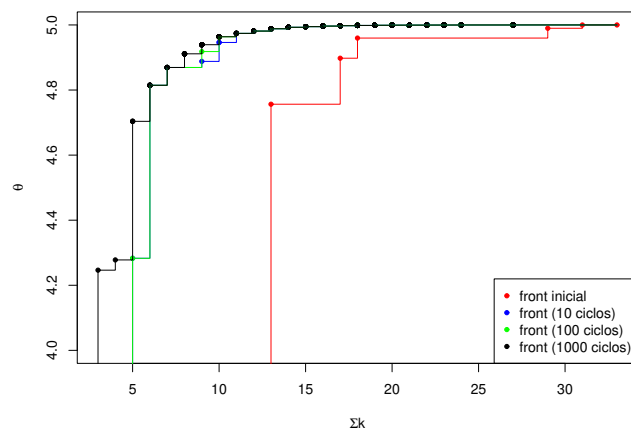
Finalmente, nas topologias divisão e fusão, há a indicação de alguma economia de escala nas filas de divisão e de fusão, principalmente nos sistemas hipereponenciais ($s^2 = 1, 5$). De fato, nas topologias divisão e fusão, a alocação em alguns casos não alcança a soma das alocações nas filas pós divisão e pré fusão, respectivamente. Também esse é um resultado bastante animador, pois está de acordo com resultados já demonstrados na literatura da área.



(a) topologia série



(b) topologia divisão



(c) topologia fusão

Figura 6: Evolução das fronteiras de Pareto em função no número de ciclos

Tabela 1: Algumas soluções dominantes em função da topologia e do s^2

Topologia	s^2	K_1	K_2	K_3	$\sum_i K_i$	Θ
Série	0,5	1	1	1	3	3,6275
		3	3	4	10	4,7702
		7	6	7	20	4,9852
		10	10	10	30	4,9992
		11	12	12	35	4,9998
		12	12	13	37	4,9999
		13	13	15	41	5,0000
	1,5	1	1	1	3	3,2895
		3	3	4	10	4,5098
		6	7	7	20	4,9253
		10	10	10	30	4,9897
		12	12	11	35	4,9960
		14	13	13	40	4,9985
		15	15	15	45	4,9995
		18	16	16	50	4,9998
		19	19	20	58	5,0000
Divisão	0,5	1	1	1	3	4,2469
		5	3	2	10	4,9635
		10	6	4	20	4,9994
		11	6	5	22	4,9997
		12	6	5	23	4,9998
		12	7	5	24	4,9999
		13	8	6	27	5,0000
	1,5	1	1	1	3	4,0063
		5	3	2	10	4,8916
		10	6	4	20	4,9945
		12	6	4	22	4,9969
		12	7	4	23	4,9976
		12	7	5	24	4,9983
		14	8	5	27	4,9993
		16	8	6	30	4,9997
		19	10	7	36	5,0000
Fusão	0,5	1	1	1	3	4,2464
		2	3	5	10	4,9637
		4	6	10	20	4,9994
		5	7	10	22	4,9997
		5	7	11	23	4,9998
		5	7	12	24	4,9999
		6	8	13	27	5,0000
	1,5	1	1	1	3	4,0039
		2	3	5	10	4,8925
		4	6	10	20	4,9945
		4	6	12	22	4,9969
		5	6	12	23	4,9976
		5	7	12	24	4,9983
		5	8	14	27	4,9993
		7	11	18	36	5,0000

4 CONCLUSÕES E OBSERVAÇÕES FINAIS

Neste trabalho, um algoritmo de enxame de partículas foi analisado para resolução de um problema de otimização em redes de filas. O *throughput* de uma rede de filas acíclica, finita e de tempo de serviço geral, de servidor único, foi maximizada e a capacidade total alocada foi minimizada.

Um algoritmo de otimização de enxame de partículas multiobjetivo foi descrito e empregado. Fronteiras de Pareto aproximadas foram obtidas. Nesta abordagem, recorreu-se à utilização de uma ferramenta clássica para avaliar aproximadamente o desempenho de redes de filas finitas, nomeadamente o método de expansão generalizado (GEM), obtendo-se soluções Pareto sub-ótimas.

Com relação ao tempo de processamento, a nova abordagem parece ser eficiente, com tempos de execução em torno de 2 minutos. Assim, experimentos futuros podem ser realizados para avaliar a flexibilidade desta abordagem em relação ao número de iterações e tamanhos de enxames, seu uso em diferentes redes de filas, com diferentes configurações (por exemplo, série, fusão, divisão e mistas), e com diferentes números de nós, de taxas de chegada e de variabilidades do tempo de serviço.

Investigações futuras devem ser executadas para determinar a aplicabilidade desta abordagem na determinação de condições sub-ótimas em outras redes de filas. Por exemplo, este método pode ser aplicado para otimizar redes gerais de filas finitas com múltiplos servidores. Além disso, pesquisas futuras devem ser conduzidas para avaliar os algoritmos em situações da vida real.

REFERÊNCIAS

- AZIMI, P.; ASADOLLAHI, A. Developing a new bi-objective functions model for a hierarchical location-allocation problem using the queuing theory and mathematical programming. *Journal of Optimization in Industrial Engineering*, QIAU, v. 12, n. 2, p. 149–154, 2019.
- COELLO-COELLO, C. A.; LECHUGA, M. S. MOPSO: A proposal for multiple objective particle swarm optimization. In: *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No.02TH8600)*. [S.l.: s.n.], 2002. v. 2, p. 1051–1056.
- 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.
- CRUZ, F. R. B.; DUARTE, A. R.; WOENSEL, T. van. Buffer allocation in general single-server queueing networks. *Computers & Operations Research*, v. 35, n. 11, p. 3581–3598, 2008.
- CRUZ, F. R. B. et al. Throughput maximization of queueing networks with simultaneous minimization of service rates and buffers. *Mathematical Problems in Engineering*, v. 2012, n. Article ID 692593, p. 19, 2012.
- DEB, K. et al. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, v. 6, n. 2, p. 182–197, 2002.
- DENG, W. et al. Study on an improved adaptive PSO algorithm for solving multi-objective gate assignment. *Applied Soft Computing*, Elsevier, v. 59, p. 288–302, 2017.
- FAN, Z. et al. An improved multiobjective particle swarm optimization algorithm using minimum distance of point to line. *Shock and Vibration*, v. 2017, p. 1–16, 2017.
- GROSS, D. et al. *Fundamentals of queueing theory*. 4th edition. ed. New York, NY: Wiley - Interscience, 2009.
- HAJIPOUR, V.; PASANDIDEH, S. H. R. Proposing an adaptive particle swarm optimization for a novel bi-objective queueing facility location model. *Economic Computation and Economic Cybernetics Studies and Research*, v. 46, n. 3, p. 223–240, 2012.
- JIA, C.; ZHU, H. An improved multiobjective particle swarm optimization based on culture algorithms. *Algorithms*, v. 10, n. 2, p. 46, 2017.
- 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.
- KENNEDY, J.; EBERHART, R. Particle swarm optimization. In: *Proceedings of ICNN'95 - International Conference on Neural Networks*. [S.l.: s.n.], 1995. v. 4, p. 1942–1948.

- KERBACHE, L.; SMITH, J. M. The generalized expansion method for open finite queueing networks. *European Journal of Operational Research*, v. 32, p. 448–461, 1987.
- KIMURA, T. A transform-free approximation for the finite capacity $M/G/s$ queue. *Operations Research*, v. 44, n. 6, p. 984–988, 1996.
- LEONG, W.; YEN, G. G. PSO-based multiobjective optimization with dynamic population size and adaptive local archives. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, IEEE, v. 38, n. 5, p. 1270–1293, 2008.
- MARTINS, H. S. R. et al. Modeling and optimization of buffers and servers in finite queueing networks. *OPSEARCH*, Springer, v. 56, n. 1, p. 123–150, 2019.
- MENASCÉ, D. A. QoS issues in web services. *IEEE Internet Computing*, v. 6, n. 6, p. 72–75, 2002.
- MOREIRA, G.; PAQUETE, L. Guiding under uniformity measure in the decision space. In: *Proceedings of the 2019 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*. [S.l.: s.n.], 2019. p. 1–6.
- OLIVEIRA, D. R. X. et al. Spatial cluster analysis using particle swarm optimization and dispersion function. *Communications in Statistics - Simulation and Computation*, Taylor & Francis, p. 1–18, 2019.
- OSORIO, C.; BIERLAIRE, M. An analytic finite capacity queueing network model capturing the propagation of congestion and blocking. *European Journal of Operational Research*, v. 196, n. 3, p. 996–1007, 2009.
- SHARAFI, M.; ELMEKKAWY, T. Y. Multi-objective optimal design of hybrid renewable energy systems using PSO-simulation based approach. *Renewable Energy*, v. 68, p. 67–79, 2014. ISSN 0960-1481.
- 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.
- 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.
- SMITH, J. M.; CRUZ, F. R. B.; WOENSEL, T. van. Topological network design of general, finite, multi-server queueing networks. *European Journal of Operational Research*, v. 201, n. 2, p. 427–441, 2010.
- SOUZA, G. L. et al. A novel formulation for multi-objective optimization of general finite single-server queueing networks. In: *2020 IEEE Congress on Evolutionary Computation (CEC)*. [S.l.: s.n.], 2020. p. 1–8.
- SOUZA, G. L. et al. Post-processing improvements in multi-objective optimization of general single-server finite queueing networks. *IEEE Latin America Transactions*, v. 21, n. 3, p. 381–388, 2023.

- TRIVEDI, V.; VARSHNEY, P.; RAMTEKE, M. A simplified multi-objective particle swarm optimization algorithm. *Swarm Intelligence (in press)*, p. 1–34, 2019. ISSN 1935-3820.
- ZHAO, X. et al. An improved mixed-integer multi-objective particle swarm optimization and its application in antenna array design. In: *2013 5th IEEE International Symposium on Microwave, Antenna, Propagation and EMC Technologies for Wireless Communications*. [S.l.: s.n.], 2013. p. 412–415.