

## DO CAOS À EFICIÊNCIA: SOLUÇÕES PARA TRANSFORMAR FILAS EM VANTAGEM COMPETITIVA PARA PEQUENOS NEGÓCIOS

FROM CHAOS TO EFFICIENCY: SOLUTIONS TO TRANSFORM QUEUES INTO A COMPETITIVE ADVANTAGE FOR SMALL BUSINESSES

DEL CAOS A LA EFICIENCIA: SOLUCIONES PARA TRANSFORMAR LAS COLAS EN UNA VENTAJA COMPETITIVA PARA LOS PEQUEÑOS NEGOCIOS

Kennedy Maxwell Oliveira Seixas<sup>1</sup>, Antonio Geraldo Harb<sup>2</sup>

DOI: 10.54899/dcs.v23i93.6490

Recibido: 01/07/2026 | Aceptado: 27/07/2026 | Publicación en línea: 05/08/2026.

### RESUMO

O gerenciamento de filas é uma tarefa crítica para pequenas empresas, particularmente nas indústrias alimentícia e de varejo, onde a flutuação na demanda pode afetar a eficiência e a experiência do cliente. Nesse contexto, esta pesquisa foi projetada para enfrentar os gargalos dos negócios e propor soluções para superá-los, de modo que os tempos de espera possam ser transformados em vantagem competitiva. Assim, o objetivo deste estudo é analisar o gerenciamento de filas em pequenas empresas em Manaus por meio da Teoria das Filas e da Simulação de Monte Carlo. A pesquisa é conduzida em termos de dados, descritiva e qualitativo-quantitativa e é baseada em observação não participante em dois estabelecimentos comerciais. Os resultados mostraram que a relação entre demanda e capacidade de serviço tem um efeito direto no tempo de espera e na saturação do sistema. A simulação de Monte Carlo também permitiu uma análise aprimorada dos dados observados e conclusões mais robustas para diferentes cenários operacionais. Concluímos que os objetivos foram alcançados e que o uso integrado da Teoria das Filas e da simulação é uma alternativa viável para decisões gerenciais e melhorias operacionais para pequenas empresas.

**Palavras-chave:** Teoria das Filas. Simulação de Monte Carlo. Gestão de Filas. Pequenos Negócios. Vantagem Competitiva.

### ABSTRACT

Queue management is a critical challenge for small businesses, particularly in the food service and retail sectors, where fluctuations in customer demand can affect operational efficiency and customer experience. In this context, this study was designed to identify operational bottlenecks and propose solutions capable of transforming waiting time into a source of competitive advantage. Therefore, the objective of this research was to analyze queue management in small businesses located in Manaus using Queueing Theory and Monte Carlo Simulation. The research

<sup>1</sup> Graduanda em Engenharia de Produção pela Universidade do Estado do Amazonas (UEA), Manaus, Amazonas, Brasil. E-mail: kmose.eng24@uea.edu.br

<sup>2</sup> Doutor em Engenharia de Produção pela Universidade do Estado do Amazonas (UEA), Manaus, Amazonas, Brasil. E-mail: aharb@uea.edu.br Orcid: <https://orcid.org/0000-0002-6122-3123>

is characterized as applied, descriptive, and based on a qualitative-quantitative approach, using direct non-participant observation in two commercial establishments. The results showed that the relationship between customer demand and service capacity has a direct impact on waiting time and system saturation. Furthermore, Monte Carlo Simulation enabled a more comprehensive analysis of the observed data, providing more robust estimates for different operational scenarios. It is concluded that the proposed objectives were achieved and that the integrated use of Queueing Theory and simulation represents a feasible approach to support managerial decision-making and improve operational performance in small businesses.

**Keywords:** Queueing Theory. Monte Carlo Simulation. Queue Management. Small Businesses. Competitive Advantage.

## RESUMEN

La gestión de colas representa un desafío crítico para los pequeños negocios, especialmente en los sectores de alimentación y comercio minorista, donde las fluctuaciones en la demanda pueden afectar la eficiencia operativa y la experiencia del cliente. En este contexto, esta investigación fue desarrollada con el propósito de identificar los principales cuellos de botella operativos y proponer soluciones capaces de transformar el tiempo de espera en una fuente de ventaja competitiva. En este sentido, el objetivo del estudio fue analizar la gestión de colas en pequeños negocios ubicados en la ciudad de Manaus mediante la aplicación de la Teoría de Colas y la Simulación de Monte Carlo. La investigación se caracteriza por ser aplicada, descriptiva y de enfoque cualitativo-cuantitativo, basada en la observación directa no participante en dos establecimientos comerciales. Los resultados demostraron que la relación entre la demanda y la capacidad de atención influye directamente en el tiempo de espera y en el riesgo de saturación del sistema. Asimismo, la Simulación de Monte Carlo permitió ampliar el análisis de los datos observados, proporcionando estimaciones más robustas para distintos escenarios operativos. Se concluye que los objetivos propuestos fueron alcanzados y que el uso integrado de la Teoría de Colas y la simulación constituye una alternativa viable para apoyar la toma de decisiones gerenciales y promover mejoras operativas en los pequeños negocios.

**Palabras clave:** Teoría de Colas. Simulación de Monte Carlo. Gestión de Colas. Pequeños Negocios. Ventaja Competitiva.



Esta obra está bajo una [Licencia Creative Commons Atribución- NoComercial 4.0 Internacional](https://creativecommons.org/licenses/by-nc/4.0/)

---

## INTRODUÇÃO

Filas são um problema comum enfrentado no serviço presencial e são mais do que uma coleção de clientes esperando por serviços: elas significam a relação entre demanda, capacidade de serviço e a organização do fluxo de operação em pequenas empresas. Quando o número de clientes que chegam ao sistema temporariamente excede a capacidade de serviço do sistema, eles

atrasam, congestionam e a fluidez do serviço diminui. Por exemplo, uma pesquisa sobre teoria das filas mostra que o tempo de espera é um fator chave que afeta a qualidade do serviço e que o tamanho dos atendentes pode levar a muito tempo de espera e pouco trabalho (Figueiredo; Rocha, 2010). No setor de serviços, a espera também afeta a percepção do cliente.

A gestão de filas é vista como algo que afeta a qualidade percebida e a modelagem do sistema pode ajudar os gestores a entender o que impulsiona a formação de filas e a melhoria dos processos (Schons; Rados, 2009). E no varejo isso é mais claro porque tempo e produtividade estão diretamente relacionados à satisfação dos clientes e à competitividade dos negócios (Oliveira *et al.*, 2021). Este problema é ainda mais agudo em pequenos negócios em Manaus, pois enfrentam o problema de recursos limitados, falta de espaço e alta dependência de horas de pico. Além disso, fatores ambientais típicos da Amazônia, como calor, umidade e desconforto em ambientes mal ventilados, podem agravar a sensação de espera. Assim, a fila deixa de ser apenas um problema operacional e também é um custo de oportunidade para os clientes desistirem de comprar, não voltarem ou associarem o local como desorganizado.

Diante desse contexto, esta pesquisa apresenta a seguinte situação-problema: **como a aplicação da Teoria das Filas, associada à modelagem operacional e à Simulação de Monte Carlo, pode contribuir para reduzir gargalos de atendimento e transformar a gestão da espera em vantagem competitiva para pequenos negócios em Manaus?**

A partir dessa questão, o objetivo geral do estudo é analisar a gestão de filas em pequenos negócios do setor de alimentação e varejo, utilizando a Teoria das Filas e a Simulação de Monte Carlo para identificar gargalos operacionais, estimar riscos de saturação e propor melhorias no atendimento. Para isso, busca-se mapear o fluxo de chegada, espera, atendimento e saída dos clientes; calcular indicadores como taxa de chegada, taxa de atendimento, tempo médio de espera e taxa de ocupação do sistema; comparar o desempenho dos estabelecimentos em períodos de pico e de menor movimento; simular cenários probabilísticos de funcionamento; e propor alternativas operacionais de baixo custo voltadas à redução da espera e à melhoria da experiência do cliente.

A pesquisa parte de duas hipóteses. **H1: em pequenos negócios presenciais, o aumento da taxa de ocupação do sistema está associado ao crescimento do tempo médio de espera e do tamanho médio da fila, especialmente em horários de pico. H2: no contexto amazônico, condições ambientais e limitações de layout reduzem a tolerância do cliente à espera, aumentando a probabilidade de abandono ou recusa de ingresso na fila.** Dessa forma, o

trabalho visa estabelecer que a gestão de filas, se respaldada por dados, modelagem e análise crítica, pode não apenas ser uma resposta emergencial à congestão, mas também uma ferramenta de melhoria operacional e de diferenciação competitiva.

## REFERENCIAL TEÓRICO

### Teoria das Filas e Desempenho Operacional em Serviços

A Teoria das Filas é uma das áreas clássicas da pesquisa operacional para sistemas onde as unidades demandantes estão aguardando serviço de recursos limitados. Em serviços presenciais, com essa abordagem, entendemos a relação de chegadas aleatórias, tempos de serviço e número de servidores com a disciplina de fila para o desempenho do sistema. Como a fila é mais do que uma representação visual do número de clientes se acumulando no sistema, ela pode ser vista como uma medida da relação demanda/capacidade operacional.

Em sistemas de serviço, a taxa média de chegada dos clientes é dada por  $\lambda$ , e a taxa média de serviço é dada por  $\mu$ . Se houver apenas um servidor, então a razão entre essas variáveis é a taxa de ocupação  $\rho = \lambda/\mu$ . Em sistemas com múltiplos servidores, a capacidade total de serviço será escolhida em função do número de canais disponíveis, então  $\rho = \lambda/(s\mu)$ . Isso é fundamental, pois indica a utilização do sistema e nos permite avaliar se a operação funciona com folga, estabilidade ou saturação. Os modelos clássicos de fila, como M/M/1 e M/M/s, são amplamente utilizados para estimar tempos médios de espera, tamanho médio da fila e probabilidade de serviço com espera, especialmente em sistemas sob aleatoriedade de chegadas e serviços. O aspecto chave da Teoria das Filas é que o crescimento da espera não ocorre de forma linear. À medida que a taxa de ocupação se aproxima de 1, pequenas variações na chegada de clientes ou no tempo de atendimento podem produzir aumentos expressivos no tempo médio de espera. Isso significa que um estabelecimento pode aparentar funcionamento normal em períodos de baixa demanda, mas tornar-se instável em horários de pico. Em pequenos negócios, esse comportamento é particularmente relevante porque a operação costuma contar com poucos atendentes, espaço físico restrito e baixa capacidade de absorver variações bruscas na demanda.

A Lei de Little reforça essa interpretação ao estabelecer a relação  $L = \lambda W$ , em que  $L$  representa o número médio de clientes no sistema,  $\lambda$  representa a taxa média de chegada e  $W$  representa o tempo médio de permanência. Essa relação demonstra que, quando o tempo de

permanência aumenta, a quantidade média de clientes no sistema também tende a crescer. Little (1961) formalizou essa relação, que se tornou uma das bases da análise de sistemas de filas. Para pequenos negócios, isso significa que reduzir o tempo de espera não melhora apenas a experiência do cliente, mas também reduz a ocupação do espaço, congestionamento e perda potencial de vendas.

### **Modelos M/M/1, M/M/s e Comportamento de Saturação**

Os modelos M/M/1 e M/M/s são amplamente utilizados na análise de sistemas de atendimento com chegadas aleatórias e tempos de serviço variáveis. Na notação de Kendall, o primeiro termo indica o padrão das chegadas, o segundo representa o padrão dos tempos de serviço e o terceiro informa o número de servidores. Dessa forma, o modelo M/M/1 representa um sistema com chegadas do tipo Poisson, tempos de serviço exponenciais e apenas um servidor; já o modelo M/M/s representa sistemas com múltiplos servidores em paralelo.

No modelo M/M/1, a estabilidade do sistema depende da condição  $\rho < 1$ . Quando  $\rho \geq 1$ , a taxa média de chegada é igual ou superior à capacidade média de atendimento, o que impede a existência de médias estacionárias finitas para tempo de espera e tamanho da fila. Em termos práticos, isso significa que o sistema não consegue absorver a demanda de forma estável, tornando-se vulnerável à formação contínua de filas e ao aumento do tempo de permanência dos clientes.

Em sistemas M/M/s, a presença de mais de um servidor amplia a capacidade total de atendimento e reduz a probabilidade de espera. No entanto, isso não elimina automaticamente os problemas operacionais. Mesmo com múltiplos atendentes ou caixas, filas podem continuar ocorrendo quando há concentração de demanda, variabilidade no serviço, falhas de *layout* ou etapas mal distribuídas no processo. Assim, a análise não deve se limitar ao número de servidores, mas considerar a fluidez do atendimento como um todo.

Para esta pesquisa, esses modelos são relevantes porque os estabelecimentos analisados possuem estruturas distintas de atendimento. O Estabelecimento A apresenta configuração mais próxima de um sistema de servidor único, enquanto o Estabelecimento B possui maior capacidade de atendimento em determinados pontos do processo. Essa diferença permite comparar como a taxa de ocupação, o número de servidores e a organização do fluxo afetam o desempenho operacional e a experiência de espera.

## **Simulação de Monte Carlo Aplicada à Análise de Filas**

A Simulação de Monte Carlo é uma técnica utilizada para analisar sistemas sujeitos à incerteza por meio da geração repetida de cenários probabilísticos. Em vez de trabalhar apenas com um valor médio observado, a simulação permite testar milhares de combinações possíveis de chegada, atendimento e capacidade. Rubinstein e Kroese (2016) destacam que o método de Monte Carlo é amplamente empregado em problemas estocásticos, especialmente quando há variabilidade e incerteza no comportamento do sistema.

Na análise de filas, essa técnica é útil porque serviços presenciais raramente apresentam comportamento constante. A quantidade de clientes pode variar conforme horário, dia da semana, clima, eventos locais e hábitos de consumo. Da mesma forma, o tempo de atendimento pode mudar de acordo com o tipo de pedido, forma de pagamento, experiência do atendente e interrupções no processo. Por isso, analisar apenas médias observadas pode esconder situações críticas que aparecem em cenários de maior demanda ou maior variabilidade.

Nesta pesquisa, os dados reais coletados em campo são utilizados como parâmetros de entrada da simulação. A taxa média de chegada ( $\lambda$ ) alimenta a distribuição de Poisson, utilizada para representar a quantidade de clientes que chegam em determinado intervalo de tempo. A taxa média de atendimento ( $\mu$ ) alimenta a distribuição exponencial, utilizada para representar a variabilidade dos tempos de serviço. Essa estrutura é coerente com os modelos M/M/1 e M/M/s, nos quais as chegadas e os atendimentos são tratados como processos probabilísticos.

A simulação não substitui a coleta real, mas amplia sua capacidade analítica. Os dois dias observados fornecem a base empírica do estudo; a Simulação de Monte Carlo permite projetar como o sistema poderia se comportar em múltiplos dias virtuais, sob diferentes combinações de demanda e atendimento. Dessa forma, a limitação temporal da coleta não é escondida, mas tratada metodologicamente. Isso fortalece o trabalho porque permite estimar risco de saturação, comportamento em horários críticos e impacto potencial de intervenções operacionais.

## **Espera, Servicescape e Vantagem Competitiva em Pequenos Negócios**

A espera em serviços não é avaliada apenas pelo tempo cronológico. Dois clientes podem esperar o mesmo número de minutos e, ainda assim, perceber a experiência de formas diferentes. Essa percepção depende de fatores como conforto, clareza do processo, sensação de justiça,

previsibilidade e organização do espaço. Portanto, a gestão de filas deve considerar tanto os indicadores matemáticos do sistema quanto a experiência percebida pelo consumidor.

O conceito de *servicescape*, proposto por Bitner (1992), destaca que o ambiente físico influencia as respostas cognitivas, emocionais e comportamentais de clientes e funcionários. Elementos como *layout*, sinalização, temperatura, iluminação, circulação e densidade de pessoas afetam a forma como o cliente percebe o serviço. Em uma fila, isso significa que um ambiente desorganizado, quente ou confuso pode tornar a espera mais desconfortável, mesmo quando o tempo objetivo não é extremamente elevado.

Em pequenos negócios, a fila também representa custo de oportunidade. Cada cliente que abandona a fila ou decide não entrar no estabelecimento representa uma venda perdida. Além disso, clientes aguardando por muito tempo ocupam espaço, dificultam a circulação e podem transmitir uma imagem de desorganização. Por isso, reduzir filas não é apenas uma melhoria operacional, mas uma forma de proteger receita, melhorar a percepção de qualidade e aumentar a chance de retorno do cliente.

No contexto amazônico, essa discussão ganha ainda mais importância. Condições como calor, umidade, ventilação limitada e espaços reduzidos podem aumentar o desconforto físico durante a espera, reduzindo a tolerância do consumidor. Dessa forma, intervenções simples, como reorganização do *layout*, separação de fluxos, sinalização clara, melhoria da ventilação e ativação de capacidade extra em horários críticos, podem transformar a fila de um problema recorrente em um diferencial competitivo. Para pequenos negócios, a eficiência no atendimento pode funcionar como vantagem competitiva justamente porque melhora a experiência sem exigir grandes investimentos.

## METODOLOGIA

Esta pesquisa caracteriza-se como aplicada, descritiva e quali-quantitativa. É aplicada porque utiliza conceitos da Teoria das Filas e da Simulação de Monte Carlo para analisar um problema real de atendimento em pequenos negócios. É descritiva porque busca registrar e interpretar o comportamento das filas observadas, considerando variáveis como chegada de clientes, tempo de atendimento, tempo de espera, ocupação do sistema e abandono. A abordagem quali-quantitativa justifica-se pelo uso combinado de dados numéricos e observações sobre *layout*, conforto ambiental e organização do atendimento.

Quanto aos procedimentos, trata-se de um estudo de caso múltiplo, realizado em dois pequenos estabelecimentos de serviços presenciais localizados em Manaus, identificados como Estabelecimento A e Estabelecimento B. A identificação por letras foi adotada para preservar a identidade dos empreendimentos. Essa escolha metodológica permite comparar realidades operacionais distintas e compreender como a gestão de filas pode ser utilizada como ferramenta de melhoria da eficiência e da experiência do cliente.

## Coleta e Dados Utilizados

A coleta de dados foi realizada por meio de observação direta não participante e a observação ocorreu em dois dias de funcionamento, contemplando períodos de pico e de menor movimento, com o objetivo de registrar diferentes condições de demanda e capacidade de atendimento.

Durante a coleta, foram registrados o número de clientes observados, a quantidade de atendentes disponíveis, o tempo médio de atendimento, o tempo médio de espera em fila, o maior tamanho de fila observado e as ocorrências de abandono ou recusa de ingresso na fila. Também foi considerada a duração de cada período observado, pois esse dado é necessário para estimar a taxa média de chegada dos clientes ao sistema.

Tabela 1. Dados operacionais observados.

| Estabelecimento | Dia | Duração da observação | Período observado | Nº de atendentes | Clientes observados | Tempo médio de atendimento | Tempo médio de espera | Maior fila observada | Abandonos |
|-----------------|-----|-----------------------|-------------------|------------------|---------------------|----------------------------|-----------------------|----------------------|-----------|
| A               | 1   | 180 min               | Pico              | 1                | 35                  | 4,2 minutos                | 5 minutos             | 7                    | 3         |
| A               | 1   | 90 min                | Menor movimento   | 1                | 16                  | 3,5 minutos                | 3 minutos             | 3                    | 0         |
| A               | 2   | 180 min               | Pico              | 1                | 40                  | 4,1 minutos                | 5,6 minutos           | 8                    | 1         |
| A               | 2   | 90 min                | Menor movimento   | 1                | 22                  | 2,7 minutos                | 3 minutos             | 3                    | 0         |
| B               | 1   | 90 min                | Pico              | 2                | 55                  | 2,5 minutos                | 4 minutos             | 12                   | 0         |
| B               | 1   | 90 min                | Menor movimento   | 2                | 30                  | 2,6 minutos                | 2 minutos             | 5                    | 0         |
| B               | 2   | 90 min                | Pico              | 2                | 46                  | 2,5 minutos                | 3,6 minutos           | 10                   | 0         |
| B               | 2   | 90 min                | Menor movimento   | 2                | 34                  | 2 minutos                  | 2 minutos             | 4                    | 0         |

Fonte: Elaborado pelos autores.

Os dados foram organizados em planilha eletrônica e serviram como base para o cálculo das variáveis utilizadas na análise de filas. A taxa média de chegada dos clientes ( $\lambda$ ) foi obtida

pela razão entre o número de clientes observados e a duração da observação. A taxa média de atendimento ( $\mu$ ) foi calculada a partir do inverso do tempo médio de atendimento, enquanto a taxa de ocupação ( $\rho$ ) considerou a relação entre demanda, capacidade de atendimento e número de servidores disponíveis. Os indicadores de tempo médio de espera, maior fila observada e abandonos foram mantidos como variáveis de comparação e interpretação dos cenários simulados.

### Modelagem dos Parâmetros e Simulação de Monte Carlo

A análise tomou como referência os modelos M/M/1 e M/M/s, conforme a estrutura observada em cada estabelecimento. O modelo M/M/1 foi utilizado para representar situações com atendimento concentrado em um único servidor, enquanto o modelo M/M/s foi considerado para sistemas com múltiplos canais de atendimento. A disciplina de fila adotada foi FIFO (*First In, First Out*), ou seja, o primeiro cliente a chegar é o primeiro a ser atendido, exceto em situações específicas de atendimento preferencial observadas em campo.

A partir dos dados observados, foram calculadas três variáveis principais: taxa média de chegada dos clientes ( $\lambda$ ), taxa média de atendimento ( $\mu$ ) e taxa de ocupação do sistema ( $\rho$ ). Nos casos com apenas um servidor, utilizou-se ( $\rho = \lambda/\mu$ ). Nos casos com mais de um servidor, utilizou-se ( $\rho = \lambda/(s\mu)$ ), em que (s) representa o número de atendentes disponíveis.

Tabela 2. Parâmetros calculados para análise de filas.

| Estabelecimento | Dia | Período         | $\lambda$<br>clientes/min | $\mu$<br>atendimentos/min | $\rho$ |
|-----------------|-----|-----------------|---------------------------|---------------------------|--------|
| A               | 1   | Pico            | 0,194                     | 0,238                     | 0,817  |
| A               | 1   | Menor movimento | 0,178                     | 0,286                     | 0,622  |
| A               | 2   | Pico            | 0,222                     | 0,244                     | 0,911  |
| A               | 2   | Menor movimento | 0,244                     | 0,37                      | 0,66   |
| B               | 1   | Pico            | 0,611                     | 0,4                       | 0,764  |
| B               | 1   | Menor movimento | 0,333                     | 0,385                     | 0,433  |
| B               | 2   | Pico            | 0,511                     | 0,4                       | 0,639  |
| B               | 2   | Menor movimento | 0,378                     | 0,5                       | 0,378  |

Fonte: Elaborado pelos autores.

Devido à limitação temporal da coleta, os parâmetros observados foram utilizados como entrada para uma Simulação de Monte Carlo. A taxa de chegada ( $\lambda$ ) foi utilizada para representar

a chegada aleatória de clientes por meio da distribuição de Poisson, enquanto o tempo médio de atendimento foi utilizado para gerar tempos de serviço com base na distribuição exponencial. A simulação foi executada com 5.000 repetições para cada cenário observado, reproduzindo virtualmente as condições de chegada e atendimento registradas na coleta. Com isso, foi possível estimar o comportamento provável das filas sob variações aleatórias de demanda e tempo de serviço.

## RESULTADOS E DISCUSSÃO

A análise dos dados observados e dos cenários simulados permitiu comparar o desempenho operacional dos dois estabelecimentos estudados. O Estabelecimento A operou com apenas um atendente em todos os períodos analisados, enquanto o Estabelecimento B contou com dois atendentes. Essa diferença de capacidade influenciou diretamente o comportamento das filas, especialmente nos horários de pico.

Nos dados observados, o Estabelecimento A apresentou os maiores sinais de criticidade nos períodos de pico. No primeiro dia, foram observados 35 clientes, tempo médio de atendimento de 4,2 minutos, espera média de 5 minutos, maior fila de 7 pessoas e 3 abandonos. No segundo dia de pico, o fluxo aumentou para 40 clientes, com tempo médio de atendimento de 4,1 minutos, espera média de 5,6 minutos, maior fila de 8 pessoas e 1 abandono. Esses dados indicam que, quando a demanda se concentra em determinados horários, a presença de apenas um atendente limita a capacidade de resposta do sistema.

No Estabelecimento B, embora o número de clientes observados tenha sido maior nos períodos de pico, o sistema apresentou melhor capacidade de absorção da demanda. No primeiro dia de pico, foram observados 55 clientes, com tempo médio de atendimento de 2,5 minutos e espera média de 4 minutos. No segundo dia de pico, foram observados 46 clientes, com o mesmo tempo médio de atendimento e espera média de 3,6 minutos. Apesar de a maior fila observada ter chegado a 12 pessoas no primeiro dia e 10 pessoas no segundo, não houve registro de abandono, o que sugere maior estabilidade operacional em comparação ao Estabelecimento A.

A Simulação de Monte Carlo reforçou essa diferença entre os estabelecimentos. A partir dos parâmetros observados, foram gerados cenários virtuais de operação, permitindo estimar o comportamento provável das filas sob variações aleatórias de chegada e atendimento. A Tabela 3 apresenta os principais resultados médios simulados.

Tabela 3. Resultados médios da Simulação de Monte Carlo.

| Estabelecimento | Dia | Período         | Clientes médios simulados | Tempo médio de espera simulado | Taxa média de ocupação simulada | Prob. ocupação acima de 85% | Prob. saturação |
|-----------------|-----|-----------------|---------------------------|--------------------------------|---------------------------------|-----------------------------|-----------------|
| A               | 1   | Pico            | 34,916                    | 10,111 min                     | 0,817                           | 40,60%                      | 16,90%          |
| A               | 1   | Menor movimento | 15,931                    | 3,950 min                      | 0,619                           | 15,30%                      | 5,80%           |
| A               | 2   | Pico            | 40,058                    | 13,546 min                     | 0,913                           | 60,80%                      | 31,80%          |
| A               | 2   | Menor movimento | 21,935                    | 3,671 min                      | 0,661                           | 16,90%                      | 5,40%           |
| B               | 1   | Pico            | 55,18                     | 2,523 min                      | 0,767                           | 27,10%                      | 5,90%           |
| B               | 1   | Menor movimento | 29,977                    | 0,498 min                      | 0,432                           | 0,10%                       | 0,00%           |
| B               | 2   | Pico            | 46,084                    | 1,411 min                      | 0,64                            | 6,90%                       | 0,70%           |
| B               | 2   | Menor movimento | 33,966                    | 0,290 min                      | 0,377                           | 0,00%                       | 0,00%           |

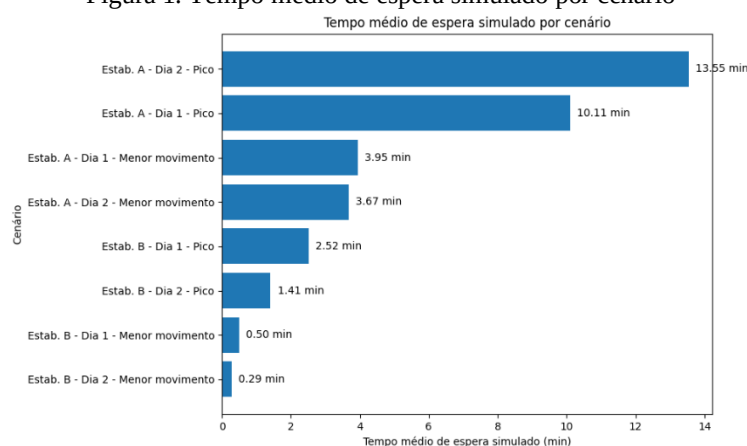
Fonte: Elaborado pelos autores.

Os resultados simulados mostram que o Estabelecimento A apresentou os maiores tempos médios de espera nos períodos de pico. No primeiro dia de pico, a espera média simulada foi de 10,111 minutos, com taxa média de ocupação de 0,817. No segundo dia de pico, a situação foi ainda mais crítica, com espera média simulada de 13,546 minutos e ocupação média de 0,913. Esse valor indica que o sistema operou muito próximo do limite de sua capacidade, aumentando a probabilidade de formação de filas longas e de abandono por parte dos clientes.

No Estabelecimento B, os tempos médios simulados foram menores, mesmo com maior quantidade de clientes. No primeiro dia de pico, o tempo médio de espera simulado foi de 2,523 minutos, com ocupação média de 0,767. No segundo dia de pico, a espera média simulada foi de 1,411 minuto, com ocupação média de 0,640. Esses resultados indicam que a presença de dois atendentes reduziu a pressão sobre o sistema, permitindo maior fluidez no atendimento.

A comparação entre os dois estabelecimentos demonstra que a formação de filas não depende apenas da quantidade de clientes, mas da relação entre demanda e capacidade de atendimento. O Estabelecimento B recebeu mais clientes em determinados períodos, porém apresentou menor tempo médio de espera simulado devido à maior capacidade de serviço. Já o Estabelecimento A, mesmo com menor volume de clientes, apresentou maior risco de congestionamento por operar com apenas um atendente e tempos médios de atendimento mais elevados.

Figura 1. Tempo médio de espera simulado por cenário



Fonte: Elaborado pelos autores.

Outro ponto relevante está relacionado ao risco de saturação. Nos cenários simulados, o Estabelecimento A apresentou maior probabilidade de ocupação acima de 85%, especialmente no segundo dia de pico, em que esse risco chegou a aproximadamente 60,8%. Nesse mesmo cenário, a probabilidade de saturação foi de aproximadamente 31,8%. Isso reforça que pequenos aumentos na chegada de clientes ou no tempo de atendimento podem gerar crescimento expressivo da fila. No Estabelecimento B, as probabilidades de saturação foram menores, principalmente nos períodos de menor movimento, indicando maior margem operacional.

Esses resultados sustentam a primeira hipótese do estudo, segundo a qual o aumento da taxa de ocupação do sistema está associado ao crescimento do tempo médio de espera e do tamanho médio da fila, especialmente em horários de pico. A hipótese é mais evidente no Estabelecimento A, onde a alta ocupação coincidiu com maior espera, maior fila observada e ocorrência de abandono ou recusa.

A segunda hipótese, relacionada à influência das condições ambientais e do layout na tolerância do cliente à espera, foi analisada de forma qualitativa. Durante a observação, verificou-se que filas mais visíveis, ambientes com circulação limitada e maior concentração de clientes tendem a tornar a espera mais desconfortável. Embora esses fatores não tenham sido mensurados numericamente, eles ajudam a explicar por que a fila pode representar não apenas um problema operacional, mas também uma ameaça à experiência do cliente e à imagem do estabelecimento.

Diante dos resultados, observa-se que intervenções simples podem contribuir para melhorar o desempenho dos pequenos negócios analisados. No Estabelecimento A, recomenda-se avaliar a ativação temporária de um segundo atendente nos horários de pico, além da reorganização da fila e da simplificação de etapas do atendimento. No Estabelecimento B, apesar

do melhor desempenho operacional, ainda há oportunidade de melhorar a experiência do cliente por meio de sinalização, organização do espaço e monitoramento dos horários de maior fluxo.

A partir desses resultados, observa-se que os objetivos específicos propostos foram contemplados ao longo da análise. O mapeamento do fluxo de clientes permitiu identificar os momentos de maior pressão sobre o atendimento; o cálculo dos indicadores de filas possibilitou avaliar a relação entre demanda e capacidade; a comparação entre períodos de pico e menor movimento evidenciou diferenças no desempenho operacional dos estabelecimentos; e a Simulação de Monte Carlo permitiu estimar o comportamento provável das filas em cenários ampliados. Com isso, os dados forneceram base para a proposição de melhorias operacionais voltadas à redução da espera e ao aumento da fluidez no atendimento.

Portanto, os resultados indicam que a gestão de filas pode ser utilizada como ferramenta de melhoria operacional e vantagem competitiva. Ao transformar dados simples de observação em indicadores de desempenho, os gestores conseguem identificar gargalos, estimar riscos de saturação e propor ajustes de baixo custo. A Simulação de Monte Carlo ampliou essa análise ao projetar cenários prováveis de funcionamento, permitindo avaliar não apenas o comportamento observado, mas também o risco operacional sob diferentes condições de demanda e atendimento.

## **CONSIDERAÇÕES FINAIS**

Esta pesquisa teve como objetivo analisar a gestão de filas em pequenos negócios de Manaus, utilizando a Teoria das Filas e a Simulação de Monte Carlo como ferramentas de apoio à identificação de gargalos operacionais, estimativa de riscos de saturação e proposição de melhorias no atendimento. Com base nos dados dos Estabelecimentos A e B, foi possível verificar que as filas não são determinadas apenas pelo número de clientes, mas também pela demanda, capacidade de atendimento, tempo médio de serviço e como o fluxo de operação é organizado.

Os dados mostraram que o Estabelecimento A era mais crítico durante os picos, pois tinha apenas um atendente e o tempo médio de serviço era maior. Quando o número de clientes era alto nesses períodos, a taxa de ocupação era suficientemente alta para levar, em muitos casos, a um tempo médio de espera e a um abandono ou recusa em entrar na fila. O Estabelecimento B, mesmo com um número maior de clientes nos horários de pico, conseguiu atender mais à demanda porque tinha dois atendentes e o tempo médio de serviço era menor. Este processo de

Simulação de Monte Carlo ajudou a estender a análise dos dados e projetou cenários prováveis em um horizonte de tempo mais longo, por exemplo, mais longo do que o período de dados que coletamos. Como resultado disso, foi possível estimar não apenas o comportamento médio das filas, mas também o risco de alta ocupação e saturação do sistema. Essa abordagem mostrou-se útil para pequenos negócios, pois transforma informações simples de observação em indicadores que podem apoiar decisões gerenciais de baixo custo.

A primeira hipótese foi sustentada pelos indicadores quantitativos, especialmente nos períodos de pico do Estabelecimento A, nos quais maiores taxas de ocupação estiveram associadas a maiores tempos médios de espera, maior fila observada e ocorrência de abandono ou recusa de ingresso. A segunda hipótese encontrou apoio nas observações qualitativas realizadas em campo, uma vez que filas visíveis, ambientes com circulação limitada e desconforto físico podem intensificar a percepção negativa da espera. No entanto, essa hipótese não foi mensurada por meio de instrumento específico de percepção do cliente, o que deve ser considerado na interpretação dos resultados.

Dessa forma, conclui-se que a gestão de filas pode deixar de ser apenas uma resposta emergencial ao congestionamento e passar a atuar como ferramenta de melhoria operacional e vantagem competitiva. Medidas simples, como ativação temporária de atendente adicional, reorganização do fluxo, melhoria da sinalização, separação de etapas do atendimento e maior atenção ao conforto do ambiente, podem reduzir a espera percebida e melhorar a experiência do cliente.

Como limitação, destaca-se que a coleta foi realizada em apenas dois dias de observação, o que restringe a generalização dos resultados. Essa limitação não é eliminada pela simulação, mas foi parcialmente mitigada pelo uso da Simulação de Monte Carlo, que permitiu projetar cenários probabilísticos a partir dos parâmetros observados. Para pesquisas futuras, recomenda-se ampliar o período de observação, incluir diferentes dias da semana, aplicar instrumentos de percepção com os clientes e considerar outros tipos de pequenos negócios, a fim de comparar padrões de fila em contextos variados de atendimento presencial.

## REFERÊNCIAS

BITNER, Mary Jo. Servicescapes: the impact of physical surroundings on customers and employees. **Journal of Marketing**, v. 56, n. 2, p. 57-71, 1992.

FIGUEIREDO, Danielle Durski; ROCHA, Silvana Heidemann. Aplicação da Teoria das Filas na otimização do número de caixas: um estudo de caso. **Iniciação Científica CESUMAR**, Maringá, v. 12, n. 2, p. 175-182, jul./dez. 2010.

LITTLE, John D. C. A proof for the queuing formula:  $L = \lambda W$ . **Operations Research**, v. 9, n. 3, p. 383-387, 1961.

OLIVEIRA, Adriano Matos de *et al.* Modelagem e análise do sistema de filas: estudo de caso em um supermercado. **Brazilian Journal of Development**, Curitiba, v. 7, n. 11, p. 101724-101745, nov. 2021. DOI: 10.34117/bjdv7n11-002.

RUBINSTEIN, Reuven Y.; KROESE, Dirk P. **Simulation and the Monte Carlo Method**. 3. ed. Hoboken: Wiley, 2016.

SCHONS, Claudio Henrique; RADOS, Gregório Varvakis. A importância da gestão de filas na prestação de serviços: um estudo na BU/UFSC. **Revista Digital de Biblioteconomia e Ciência da Informação**, Campinas, v. 6, n. 2, p. 116-135, jan./jun. 2009.