ConBRepro
XVI CONGRESSO BRASILEIRO DE ENGENHARIA DE PRODUÇÃO
EVENTO ON-LINE
02 a 04 de dezembro de 2026
Inovação e Transformação Digital como Estratégia Competitiva nas Engenharias

ANÁLISE DE SERVIÇOS DE DELIVERY A PARTIR DA TEORIA DAS FILAS E DE SIMULAÇÃO DISCRETA

Anni Pucinelli
Departamento de Engenharia de Produção e Sistemas – DEPS/UFSC
Ricardo Villarroel Dávalos
Departamento de Engenharia de Produção e Sistemas – DEPS/UFSC

Resumo: As mudanças nos hábitos de consumo ampliaram a participação do delivery nas operações de restaurantes, que passaram a conciliar diferentes canais de atendimento e recursos compartilhados. Nesse contexto, o dimensionamento da capacidade e o controle dos tempos de espera tornam-se relevantes para o desempenho operacional. Este estudo analisa uma operação de delivery de um restaurante, com o objetivo de avaliar a capacidade equivalente utilizada para representar esse fluxo e o comportamento do sistema diante de variações na demanda, produtividade e capacidade. Foram coletados dados operacionais em três campanhas independentes, considerando os horários de chegada, início do preparo e conclusão dos pedidos. A partir desses registros, foram estimados os parâmetros do sistema e adotado o modelo M/M/11 como configuração de referência. Complementarmente, foi desenvolvido um modelo de Simulação de Eventos Discretos, com 1.000 replicações independentes de Monte Carlo por cenário. O ManageQueue, aplicação computacional desenvolvida no contexto do estudo, integrou a coleta, a análise, a simulação e a apresentação dos resultados. A modelagem analítica indicou estabilidade para a configuração M/M/11, com tempo médio de espera em fila de 4,62 minutos. Na simulação, o aumento de 50% na demanda elevou a espera média para aproximadamente 26 minutos, enquanto o aumento de 20% na taxa de atendimento apresentou os menores tempos de espera entre os cenários avaliados. Os resultados demonstram que a aplicação conjunta da Teoria das Filas e da simulação permite analisar o dimensionamento da capacidade e avaliar o comportamento da operação sob diferentes condições operacionais.

Palavras-chave: Teoria das Filas, Simulação de Eventos Discretos, Delivery, Dimensionamento de Capacidade, Pesquisa Operacional.

ANALYSIS OF DELIVERY SERVICES USING QUEUEING THEORY AND DISCRETE-EVENT SIMULATION

Abstract: Changes in consumption habits have increased the relevance of delivery services in restaurant operations, which must manage different service channels and shared resources. In this context, capacity sizing and waiting-time control are relevant to operational performance. This study analyzes a restaurant delivery operation to evaluate the equivalent capacity used to represent this flow and the behavior of the system under changes in demand, productivity, and capacity. Operational data were collected through three independent campaigns, recording order arrival, preparation start, and completion times. Based on these records, the system parameters were estimated, and an M/M/11 model was adopted as the reference configuration. In addition, a Discrete-Event Simulation model was developed, with 1,000 independent Monte Carlo replications

per scenario. ManageQueue, a computational application developed within the study, integrated data collection, analysis, simulation, and presentation of results. The analytical model indicated stability for the M/M/11 configuration, with an average queue waiting time of 4.62 minutes. In the simulation, a 50% increase in demand raised the average waiting time to approximately 26 minutes, whereas a 20% increase in the service rate resulted in the lowest waiting times among the evaluated scenarios. The results demonstrate that the combined application of Queueing Theory and Discrete-Event Simulation enables capacity sizing analysis and the evaluation of operational behavior under different conditions.

Keywords: Queueing Theory, Discrete-Event Simulation, Delivery, Capacity Sizing, Operations Research.

1. Introdução

As operações de restaurantes envolvem a coordenação simultânea de demanda, recursos humanos, atividades de preparação e diferentes formas de atendimento. Em sistemas sujeitos à variabilidade, o desempenho da operação depende da relação entre a demanda e a capacidade disponível, uma vez que alterações nos processos de chegada e atendimento podem resultar na formação de filas e no aumento dos tempos de espera (HILLIER; LIEBERMAN, 2021; TAHA, 2017).

A incorporação do delivery acrescenta outro fluxo de demanda à operação do restaurante. Quando o atendimento presencial e os pedidos destinados à entrega utilizam a mesma estrutura produtiva, equipamentos e recursos humanos passam a atender demandas distintas. Nesses casos, a capacidade disponível precisa ser distribuída entre os diferentes canais, e restrições nessa capacidade podem afetar o desempenho da operação (MA; WEBB; SCHWARTZ, 2021).

Quando a demanda se aproxima da capacidade de atendimento, pequenas alterações nos processos de chegada ou de serviço podem elevar os tempos de espera. No delivery, essa relação depende não apenas dos recursos existentes no restaurante, mas da capacidade efetivamente disponível para esse fluxo, considerando seu compartilhamento com o atendimento presencial.

O presente estudo analisa uma operação real de delivery de um restaurante em que os recursos produtivos são compartilhados com o atendimento presencial. A partir desse problema, busca-se responder à seguinte questão: como o dimensionamento da capacidade equivalente utilizada para representar o fluxo de delivery e as variações nas condições de demanda, produtividade e capacidade de atendimento influenciam o desempenho do sistema de filas de uma operação de restaurante?

Assim, este artigo tem como proposta analisar o desempenho da operação por meio da aplicação conjunta da Teoria das Filas e da Simulação de Eventos Discretos, utilizando dados coletados diretamente no sistema para representar seu comportamento e avaliar diferentes condições de demanda, produtividade e capacidade de atendimento. Como suporte à execução dessas etapas, foi utilizado o ManageQueue, aplicação computacional desenvolvida no contexto do estudo e estruturada para integrar a coleta de dados, a modelagem analítica, a simulação e a apresentação dos resultados.

2. Referencial Teórico

O referencial teórico deste estudo está organizado a partir dos principais conceitos necessários à análise da operação estudada. Inicialmente, são abordadas as características dos serviços de delivery e sua relação com a gestão da capacidade em operações de serviços. Em seguida, são apresentados os fundamentos da Teoria das

Filas utilizados para representar o sistema de atendimento e, por fim, os conceitos relacionados à Simulação de Eventos Discretos, utilizada de forma complementar à modelagem analítica.

2.1 Gestão da Capacidade em Operações de Serviços

A incorporação do delivery acrescenta um fluxo de demanda às operações de restaurantes. Quando os canais presencial e de entrega compartilham a estrutura produtiva, alterações no volume de pedidos podem afetar a utilização dos recursos e o desempenho da operação. Ma, Webb e Schwartz (2021) analisam essa relação entre demanda presencial, delivery e restrições de capacidade, destacando a importância da alocação dos recursos produtivos.

A gestão da capacidade em operações de serviços está relacionada ao equilíbrio entre os recursos disponíveis e a demanda a ser atendida. Como a demanda e os tempos de processamento estão sujeitos a variações, o dimensionamento não deve considerar apenas a capacidade nominal instalada, mas também as oscilações presentes no processo de atendimento (SLACK; BRANDON-JONES; JOHNSTON, 2018). Em situações de elevada utilização, a redução da capacidade disponível para absorver essas variações pode resultar em aumento das filas e dos tempos de espera.

Green, Kolesar e Whitt (2007) destacam que o planejamento da capacidade em serviços deve considerar a variabilidade da demanda e seus efeitos sobre a necessidade de recursos. Em restaurantes com canais compartilhados, essa relação torna relevante a análise da capacidade efetivamente disponível para cada fluxo, especialmente nos períodos de maior demanda.

Estudos aplicados a serviços de alimentação também demonstram a utilização de métodos quantitativos para avaliar a relação entre demanda, capacidade e formação de filas. Guerreiro, Moraes e Dávalos (2026), ao analisarem o sistema de atendimento do Restaurante Universitário da UFSC, utilizaram Teoria das Filas e Simulação de Eventos Discretos para avaliar diferentes configurações de capacidade em uma operação de alimentação com elevada demanda.

2.2 Teoria das Filas

A Teoria das Filas permite analisar sistemas nos quais entidades chegam em busca de atendimento e podem aguardar quando os recursos disponíveis estão ocupados. Esses sistemas são caracterizados pelo processo de chegada, pelo processo de atendimento, pelo número de servidores, pela capacidade do sistema e pela disciplina da fila. A partir dessas características, podem ser obtidos indicadores relacionados à utilização dos recursos, ao tamanho das filas e aos tempos de espera (SHORTLE et al., 2018).

Entre os modelos utilizados na análise de filas estão os sistemas M/M/s, nos quais as chegadas seguem um processo de Poisson, os tempos de atendimento são exponencialmente distribuídos e existem s servidores em paralelo. Nesse modelo, λ representa a taxa média de chegada e μ a taxa média de atendimento de cada servidor. A relação entre demanda e capacidade pode ser representada pela taxa de utilização e definida na equação 1.

$$\rho = \frac{s}{\mu\lambda} \quad (1)$$

Para que um sistema M/M/s apresente equilíbrio em regime estacionário, é necessário que $\rho < 1$. Quando $\rho \geq 1$, a demanda média iguala ou supera a capacidade média de atendimento, não sendo possível estabelecer um regime estacionário com fila limitada. Mesmo quando essa condição é atendida, valores próximos da unidade indicam elevada utilização dos recursos e maior sensibilidade do sistema às variações nos processos de chegada e atendimento (KLEINROCK, 1975; SHORTLE et al., 2018).

Além da utilização, a análise pode considerar o número médio de entidades na fila (Lq), o número médio de entidades no sistema (L), o tempo médio de espera na fila (Wq) e o tempo médio de permanência no sistema (W). Em conjunto, esses indicadores permitem avaliar os efeitos de diferentes níveis de demanda e capacidade sobre o desempenho do sistema, evitando que a análise seja baseada apenas na quantidade de recursos disponíveis.

A aplicação da Teoria das Filas em problemas de dimensionamento também pode ser observada em diferentes operações de serviços. Pieper, Tancredi e Dávalos (2025), ao analisarem um terminal de contêineres, utilizaram modelos de filas para avaliar a relação entre demanda e capacidade em diferentes cenários operacionais, demonstrando a utilização da abordagem como suporte à análise de capacidade.

2.3 Simulação de Eventos Discretos

A Simulação de Eventos Discretos (SED) permite representar computacionalmente sistemas cujo estado se modifica a partir da ocorrência de eventos ao longo do tempo. Em sistemas de filas, eventos como chegadas, início de atendimento e conclusão do serviço alteram as condições do sistema, permitindo acompanhar a formação das filas, a utilização dos recursos e os tempos associados ao processo. Essa abordagem possibilita analisar sistemas dinâmicos e avaliar diferentes configurações sem realizar alterações diretamente na operação real (BANKS et al., 2010; LAW, 2015).

Em operações de serviços, a simulação pode ser utilizada para avaliar alterações na demanda, na capacidade e na organização dos recursos antes de sua implementação. Usluer (2025), por exemplo, desenvolveu um modelo de Simulação de Eventos Discretos para analisar problemas de filas em um serviço de alimentação universitário. A partir da construção de diferentes cenários, o estudo avaliou alternativas para redução das filas e identificou a realocação da capacidade como uma das estratégias de melhoria do desempenho do sistema.

A combinação entre Teoria das Filas e simulação também foi aplicada por Vieira e Dávalos (2025) na análise de um sistema de atendimento em um posto de combustível. Os autores avaliaram diferentes condições operacionais por meio da modelagem de filas e da construção de cenários simulados, utilizando os resultados para analisar alterações no desempenho do sistema.

Em modelos estocásticos, diferentes execuções de uma mesma configuração podem apresentar resultados distintos devido à aleatoriedade das variáveis consideradas. A realização de múltiplas replicações permite analisar essa variabilidade e obter estimativas mais representativas do comportamento do sistema (LAW, 2015). O método de Monte Carlo utiliza amostras aleatórias geradas de acordo com distribuições de probabilidade previamente definidas, possibilitando a realização de sucessivos experimentos e a análise estatística dos resultados obtidos (RUBINSTEIN; KROESE, 2017).

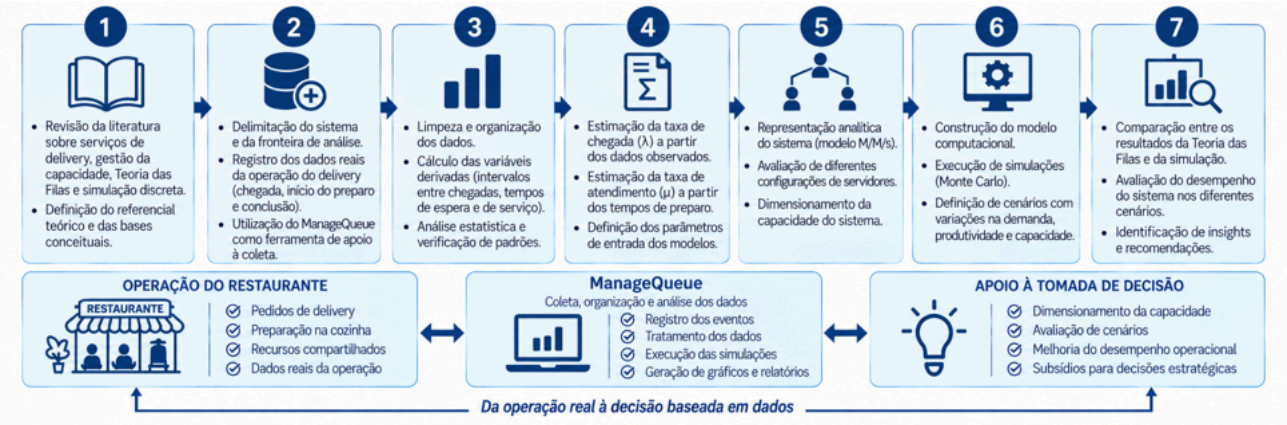
A modelagem analítica e a simulação oferecem perspectivas complementares: enquanto os modelos de filas fornecem indicadores a partir das relações entre demanda e capacidade, a SED reproduz a evolução do sistema e permite avaliar diferentes condições operacionais. Sua aplicação conjunta possibilita examinar os efeitos das alterações dos parâmetros sobre as filas, os tempos de espera e a utilização dos recursos.

3. Metodologia

O presente trabalho foi desenvolvido por meio de um estudo de caso, conduzido a partir da análise de uma operação real de delivery de um restaurante. O estudo de caso permite investigar um fenômeno dentro de seu contexto real, sendo adequado a situações em que as características do objeto analisado e do ambiente no qual ele está inserido estão diretamente relacionadas (MIGUEL, 2007).

A condução da pesquisa foi organizada em etapas sequenciais, iniciando pelo levantamento bibliográfico e pela caracterização do sistema estudado. Para apoiar a execução das etapas seguintes, foi utilizado o ManageQueue, aplicação computacional desenvolvida em Python e estruturada em quatro módulos funcionais integrados: coleta de dados, análise, simulação e geração de relatórios para apoio à tomada de decisão. O módulo de coleta permite registrar e organizar os eventos da operação; o módulo de análise realiza o tratamento dos dados, a estimação dos parâmetros e os cálculos associados à Teoria das Filas; e o módulo de simulação utiliza os parâmetros obtidos para executar os experimentos de Simulação de Eventos Discretos e as replicações de Monte Carlo. Por fim, o módulo de relatórios consolida os indicadores e resultados em gráficos, tabelas e relatórios automatizados, permitindo a comparação dos cenários. Essa estrutura integra as etapas da pesquisa em uma mesma aplicação, desde o registro dos dados até a apresentação dos resultados. A Figura 1 sintetiza as etapas metodológicas adotadas na pesquisa.

Figura 1 - Etapas metodológicas para a elaboração do estudo de caso



Fonte: Elaborado pelos autores (2026)

A pesquisa iniciou-se com uma revisão bibliográfica sobre serviços de delivery, gestão da capacidade, Teoria das Filas e Simulação de Eventos Discretos. Em seguida, a operação foi caracterizada por meio de observação direta e de informações fornecidas pelo responsável pelo estabelecimento. Os eventos associados à chegada, ao início do preparo e à conclusão dos pedidos foram registrados, organizados e tratados para obtenção das variáveis e estimação dos parâmetros utilizados nos modelos.

Na etapa de modelagem analítica, os parâmetros estimados foram utilizados para avaliar configurações com múltiplos servidores equivalentes e analisar os principais indicadores de desempenho do fluxo de delivery.

Após a modelagem analítica, o módulo de simulação do ManageQueue foi utilizado para executar o modelo de Simulação de Eventos Discretos e as replicações de Monte Carlo, incorporando a variabilidade dos processos de chegada e atendimento.

Na sequência, foram construídos cenários a partir de alterações nas condições de demanda, produtividade e capacidade de atendimento. Esses cenários permitiram avaliar o comportamento do sistema diante de diferentes condições operacionais, mantendo uma configuração de referência para comparação.

Por fim, os resultados obtidos pela Teoria das Filas e pela Simulação de Eventos Discretos foram analisados de forma conjunta. A comparação dos indicadores permitiu avaliar o comportamento do sistema na configuração de referência e nos cenários simulados, constituindo a base para a discussão dos resultados e para as recomendações apresentadas ao final do estudo.

4. Estudo de Caso

4.1 Caracterização e Delimitação do Sistema

O estudo de caso foi realizado em um restaurante que opera simultaneamente com atendimento presencial e serviço de delivery. A produção ocorre sob demanda e o cardápio é diversificado, fazendo com que diferentes pedidos utilizem recursos e etapas distintas durante o processo de preparo. Os dois canais são atendidos pela mesma estrutura produtiva, com compartilhamento de recursos ao longo da operação.

Para este estudo, a análise foi delimitada ao fluxo dos pedidos destinados ao delivery. O sistema considerado tem início no momento em que o pedido é disponibilizado para produção e termina com a conclusão de seu preparo. Dessa forma, as etapas posteriores de expedição, despacho e transporte até o cliente não fazem parte da fronteira analisada.

Cada pedido de delivery foi considerado uma entidade do sistema. Após sua entrada, o pedido pode iniciar o atendimento quando existe capacidade disponível ou permanecer em fila até a liberação de um recurso. A fila segue a disciplina FIFO (*First In, First Out*), de modo que os pedidos são encaminhados para atendimento respeitando sua ordem de chegada.

Uma característica importante da operação é a existência de diferentes estações de trabalho e recursos compartilhados entre os pedidos do salão e do delivery. Por esse motivo, o número total de colaboradores presentes no restaurante não foi considerado diretamente como o número de servidores disponíveis para o delivery. Para possibilitar a representação analítica do sistema, a capacidade utilizada para representar esse fluxo foi tratada em termos de servidores equivalentes, cujo dimensionamento foi realizado posteriormente a partir dos dados coletados e da modelagem pela Teoria das Filas.

4.2 Coleta e Tratamento dos Dados

A coleta dos dados foi realizada diretamente durante a operação do serviço de delivery, em três campanhas independentes. Para cada pedido foram registrados o horário de chegada ao sistema, o início do preparo e o horário de conclusão do atendimento. Esses

registros permitiram acompanhar o fluxo dos pedidos dentro da fronteira definida para o estudo e constituíram a base para a parametrização dos modelos.

O registro dos eventos foi realizado por meio do módulo de coleta de dados do ManageQueue, utilizado para acompanhar os pedidos e organizar os registros operacionais. A partir dos horários registrados, foram calculados os intervalos entre chegadas, os tempos de espera, os tempos de atendimento e os tempos totais de permanência no sistema. As três campanhas foram posteriormente comparadas para analisar o comportamento dos dados antes de sua utilização na modelagem.

A taxa média de chegada (λ) foi estimada a partir do fluxo de entrada dos pedidos durante o período observado, enquanto a taxa média de atendimento (μ) foi obtida a partir dos tempos de preparo registrados. Esses parâmetros foram utilizados tanto na modelagem pela Teoria das Filas quanto na Simulação de Eventos Discretos, estabelecendo uma mesma base de entrada para as duas abordagens.

4.3 Modelagem pela Teoria das Filas e Dimensionamento da Capacidade

A partir dos dados tratados, foram adotadas para a modelagem uma taxa média de chegada de $\lambda = 0{,}4600$ pedidos/min e uma taxa média de atendimento de $\mu = 0{,}0506$ pedidos/min por servidor. Como os recursos produtivos são compartilhados entre o atendimento presencial e o delivery, o número de servidores foi tratado como uma medida de capacidade equivalente utilizada para representar o fluxo analisado. Inicialmente, foi considerada a representação por um único servidor. Entretanto, para os parâmetros observados, a capacidade de atendimento de um sistema M/M/1 é inferior à taxa média de chegada, não atendendo à condição de estabilidade. Dessa forma, o dimensionamento foi realizado por meio de configurações M/M/s, aumentando-se progressivamente o número de servidores equivalentes e avaliando a relação entre demanda, capacidade e os indicadores de desempenho do sistema.

A modelagem analítica foi executada no módulo de análise do ManageQueue, responsável pelo processamento dos dados operacionais e pelo cálculo dos indicadores da Teoria das Filas. A partir das taxas estimadas, foram avaliadas diferentes configurações M/M/s, considerando a utilização dos servidores, o número médio de pedidos em fila e os tempos médios de espera. Os resultados permitiram comparar as configurações e estabelecer a capacidade equivalente utilizada como referência para a simulação.

As configurações entre M/M/1 e M/M/9 apresentaram taxa de utilização superior à unidade, indicando que a capacidade média de atendimento seria inferior à demanda observada. A primeira configuração a atender à condição de estabilidade foi o M/M/10, com taxa de utilização de aproximadamente 0,91. Apesar disso, o tempo médio de espera em fila (W_q) permaneceu em 14,92 minutos, com número médio de 6,86 pedidos aguardando atendimento.

Com a configuração M/M/11, a taxa de utilização foi reduzida para aproximadamente 0,83 e o tempo médio de espera em fila para 4,62 minutos, correspondendo a uma redução de aproximadamente 69% em relação à configuração com dez servidores. O número médio de pedidos em fila também foi reduzido para 2,13. Assim, a inclusão do décimo primeiro servidor equivalente proporcionou uma margem adicional de capacidade e uma redução expressiva dos indicadores associados à fila.

Diante desses resultados, o modelo M/M/11 foi adotado como configuração de referência para representar a capacidade equivalente do sistema de delivery. Essa configuração foi posteriormente utilizada como base para a Simulação de Eventos Discretos e para a construção dos diferentes cenários operacionais.

4.4 Simulação de Eventos Discretos e Construção dos Cenários

Após a definição do modelo M/M/11, foi desenvolvida a Simulação de Eventos Discretos utilizando como referência os mesmos parâmetros empregados na modelagem analítica. Os intervalos entre chegadas e os tempos de atendimento foram gerados a partir de distribuições exponenciais parametrizadas pelas taxas médias λ e μ, respectivamente.

A adoção das distribuições exponenciais foi realizada em conformidade com as hipóteses do modelo M/M/s utilizado como referência analítica, não constituindo, neste estudo, uma validação empírica da distribuição dos dados observados.

A simulação foi implementada no módulo específico do ManageQueue, que recebe os parâmetros definidos na etapa de análise por meio de um arquivo de configuração. O módulo utiliza esses parâmetros para gerar os tempos aleatórios de chegada e atendimento, executar as replicações de Monte Carlo e calcular os indicadores de desempenho de cada cenário. Os resultados dos experimentos são armazenados em um banco de dados SQLite e posteriormente consolidados em gráficos e relatórios, permitindo a comparação das condições operacionais avaliadas.

Na simulação, cada pedido foi representado como uma entidade que ingressa no sistema e verifica a disponibilidade dos servidores. Quando existe capacidade disponível, o atendimento é iniciado imediatamente; caso contrário, o pedido permanece em fila, seguindo a disciplina FIFO, até a liberação de um servidor. A conclusão de um atendimento libera o recurso para o próximo pedido aguardando na fila.

Os experimentos foram realizados por meio de 1.000 replicações independentes de Monte Carlo para cada cenário. A configuração base foi definida com λ = 0,460 pedidos/min, μ = 0,051 pedidos/min por servidor e 11 servidores equivalentes. A partir dessa configuração, foram construídos sete cenários adicionais para avaliar diferentes condições operacionais.

Para a análise dos resultados das replicações, foi adotado como referência um limite operacional de oito minutos para o tempo médio de espera em fila. Esse valor foi utilizado como parâmetro de comparação entre os cenários, permitindo avaliar a frequência com que o tempo médio de espera de cada replicação ultrapassou o limite adotado.

Para representar alterações na demanda, foram considerados um aumento de 20% na taxa de chegada, correspondente ao cenário de pico, e um aumento de 50%, representando uma situação de promoção. As variações na produtividade foram analisadas por meio do aumento e da redução de 20% na taxa de atendimento. Também foram avaliadas alterações na capacidade, reduzindo o número de servidores equivalentes de 11 para 10 e aumentando-o de 11 para 12. Por fim, foi considerado um

cenário crítico, combinando aumento de 30% na taxa de chegada com redução de 10% na taxa de atendimento.

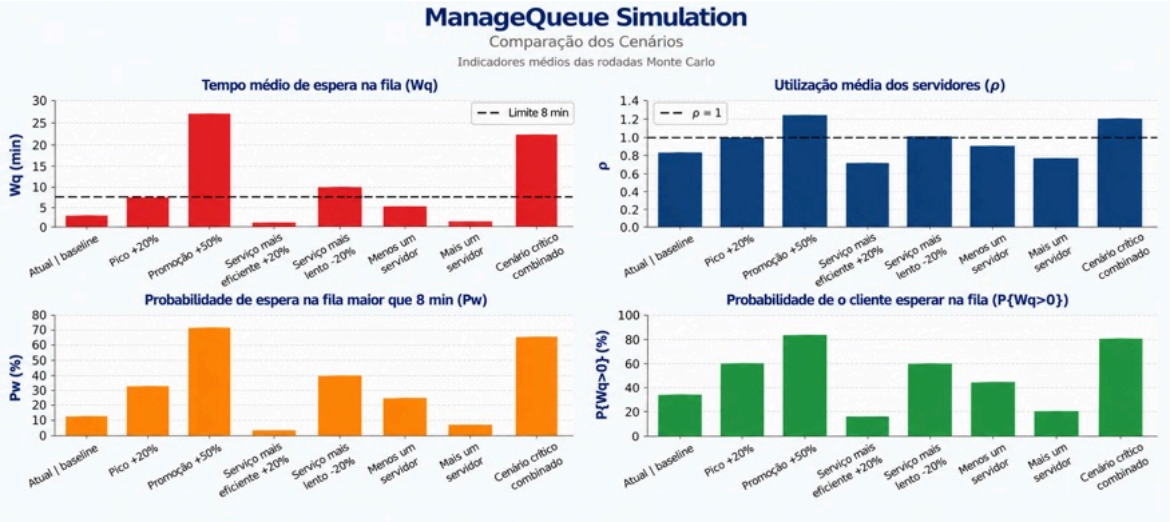
Dessa forma, foram analisados oito cenários no total, permitindo avaliar separadamente os efeitos de mudanças na demanda, na produtividade e na capacidade, além do efeito combinado de condições desfavoráveis. Os resultados das replicações foram consolidados para posterior comparação entre os cenários e com a configuração de referência.

4.5 Análise Comparativa dos Resultados

No cenário base, o sistema apresentou tempo médio de espera de aproximadamente 2,5 minutos e utilização próxima de 82%, indicando que a capacidade média de atendimento permaneceu superior à demanda observada. O aumento de 20% na demanda elevou o tempo médio de espera para aproximadamente 8 minutos e a utilização para próximo de 100%. Com o aumento de 50% na demanda, correspondente ao cenário de promoção, o tempo médio de espera atingiu aproximadamente 26 minutos e a utilização ultrapassou a unidade, indicando que a demanda média passou a superar a capacidade média de atendimento. A deterioração observada foi acompanhada pelo aumento da probabilidade de formação de filas e do risco de o tempo médio de espera exceder oito minutos.

Os cenários de produtividade apresentaram comportamento inverso. O aumento de 20% na taxa de atendimento reduziu o tempo médio de espera para aproximadamente 1 minuto e a utilização para cerca de 70%, enquanto a redução de 20% na produtividade elevou Wq para aproximadamente 10 minutos e levou a utilização para próximo da unidade. A variação do número de servidores produziu efeito semelhante: a retirada de um servidor elevou Wq para aproximadamente 5 minutos, enquanto a adição de um servidor reduziu o indicador para aproximadamente 1 minuto. Os resultados mostram, portanto, que tanto a produtividade quanto a capacidade disponível influenciam a margem do sistema para absorver variações na demanda, ilustrado na figura a seguir.

Figura 2 - Resultados Consolidados por Cenário



Fonte: Elaborado pelos Autores por meio do ManageQueue (2026)

A análise probabilística complementa os valores médios obtidos nos experimentos. No cenário base, aproximadamente 12% das replicações apresentaram tempo médio de espera superior a oito minutos. Esse percentual aumentou para cerca de 35% com o acréscimo de 20% na demanda e para aproximadamente 69% no cenário de promoção. Em sentido oposto, o aumento de 20% na taxa de atendimento reduziu esse percentual para aproximadamente 3%, enquanto a adição de um servidor resultou em valor próximo de 5%. A probabilidade de espera apresentou comportamento semelhante, sendo maior nos cenários de aumento da demanda e menor nas configurações com maior produtividade ou capacidade. Esses resultados permitem observar diferenças entre os cenários que não aparecem apenas pela comparação dos valores médios.

O cenário crítico combinado apresentou aproximadamente 23 minutos de espera média, utilização superior à unidade e cerca de 63% das replicações com tempo médio de espera superior a oito minutos. Esse comportamento evidencia a deterioração do desempenho quando o aumento da demanda ocorre simultaneamente à redução da taxa de atendimento. Em contraste, o cenário de aumento de 20% na taxa de atendimento apresentou Wq de aproximadamente 0,65 minuto, mesmo sem alteração do número de servidores.

A comparação dos cenários evidencia que o desempenho da operação depende da relação entre demanda, produtividade e capacidade equivalente. À medida que a utilização se aproxima da unidade, os tempos de espera tornam-se mais sensíveis às alterações dos parâmetros do sistema. Nos cenários em que ρ ultrapassa a unidade, a demanda média supera a capacidade média de atendimento e o modelo M/M/s deixa de apresentar equilíbrio em regime estacionário. A simulação, por trabalhar com um horizonte definido, permitiu observar o comportamento da operação também nessas condições de sobrecarga.

A modelagem pela Teoria das Filas e a simulação apresentaram resultados complementares. Para a configuração M/M/11, o modelo analítico resultou em ρ = 0,826 e Wq = 4,62 minutos, enquanto a simulação do cenário base apresentou utilização equivalente e Wq médio de 2,56 minutos. A Teoria das Filas permitiu dimensionar a capacidade equivalente e identificar a configuração de referência do sistema, enquanto a simulação possibilitou avaliar seu comportamento diante das alterações de demanda, produtividade e capacidade consideradas nos diferentes cenários.

5. Conclusões e considerações finais

Este estudo analisou uma operação de delivery com recursos compartilhados com o atendimento presencial. A aplicação conjunta da Teoria das Filas e da Simulação de Eventos Discretos permitiu dimensionar a capacidade equivalente do fluxo analisado e avaliar seu desempenho sob diferentes condições de demanda, produtividade e capacidade de atendimento.

A modelagem pela Teoria das Filas mostrou que o atendimento da condição de estabilidade não é suficiente, isoladamente, para definir uma configuração adequada. Embora o modelo M/M/10 tenha sido a primeira configuração estável, apresentou utilização de 0,908 e tempo médio de espera de 14,92 minutos. Com onze servidores

equivalentes, a utilização foi reduzida para 0,826 e o tempo médio de espera para 4,62 minutos, razão pela qual o M/M/11 foi adotado como configuração de referência para os experimentos de simulação.

Os cenários simulados mostraram que o desempenho da operação é particularmente sensível ao aumento da demanda e à redução da produtividade. O acréscimo de 50% na taxa de chegada elevou o tempo médio de espera para aproximadamente 26 minutos, enquanto o cenário crítico, combinando maior demanda e menor taxa de atendimento, também levou o sistema a uma condição de sobrecarga. Em contrapartida, o aumento de 20% na taxa de atendimento reduziu o tempo médio de espera para aproximadamente 0,65 minuto, sem alteração do número de servidores. A adição de um servidor também reduziu o tempo de espera, enquanto o cenário com aumento de 20% na taxa de atendimento apresentou a maior redução de Wq entre as condições avaliadas. Esses resultados mostram que a análise da capacidade não deve se limitar à quantidade de recursos disponíveis, mas considerar também o desempenho do processo de atendimento.

Recomenda-se que aumentos previstos na demanda, especialmente durante períodos de pico ou ações promocionais, sejam acompanhados de uma avaliação prévia da capacidade de atendimento. Os cenários analisados indicam a necessidade de considerar alternativas de reforço temporário da capacidade ou de melhoria da taxa de atendimento, observando sua viabilidade operacional e os recursos necessários à implementação.

O acompanhamento da relação entre demanda e capacidade também é relevante quando a utilização se aproxima da unidade. A redução da produtividade deve integrar o planejamento operacional, pois pode comprometer o desempenho mesmo sem aumento da demanda. Nesse sentido, a análise de cenários fornece suporte quantitativo às decisões de capacidade.

A utilização conjunta da Teoria das Filas e da simulação mostrou-se complementar ao longo da análise. Enquanto o modelo M/M/s forneceu uma referência para o dimensionamento da capacidade e para a identificação das condições de equilíbrio, a simulação permitiu avaliar a resposta do sistema às alterações de demanda, produtividade e capacidade e considerar a variabilidade entre as replicações. O ManageQueue foi utilizado como ferramenta de apoio à coleta, ao tratamento dos dados e à execução das análises, sem constituir o objeto central desta investigação.

Como limitações, o estudo representa uma operação específica e utiliza servidores equivalentes para representar de forma agregada uma estrutura em que existem recursos especializados e compartilhados entre o delivery e o atendimento presencial. Além disso, os resultados estão associados às condições observadas durante as campanhas de coleta e às hipóteses utilizadas na construção dos modelos, o que exige uma nova caracterização antes de sua aplicação a outras operações.

Como continuidade da pesquisa, recomenda-se ampliar a coleta para diferentes dias e períodos, investigar modelos que representem a utilização compartilhada e a especialização dos recursos da cozinha e incorporar variações temporais da demanda e distribuições ajustadas a conjuntos mais amplos de dados. Essas medidas podem aproximar o modelo da operação real e ampliar a avaliação das decisões de capacidade.

Referências

BANKS, Jerry; CARSON II, John S.; NELSON, Barry L.; NICOL, David M. **Discrete-event system simulation**. 5. ed. Upper Saddle River: Prentice Hall, 2010.

GREEN, Linda V.; KOLESAR, Peter J.; WHITT, Ward. Coping with time-varying demand when setting staffing requirements for a service system. **Production and Operations Management**, v. 16, n. 1, p. 13–39, 2007. DOI: 10.1111/j.1937-5956.2007.tb00164.x.

GUERREIRO, Theo Furtado; MORAES, Gustavo Alarcão Lima; DÁVALOS, Ricardo Villarroel. Simulação Discreta e Teoria das Filas aplicada a processos de atendimento: estudo de caso do Restaurante Universitário da UFSC. In: ENCONTRO NACIONAL DE ENGENHARIA DE PRODUÇÃO – ENEGEP, 46., 2026, São Paulo. **Artigo aceito para apresentação**, 2026.

HILLIER, Frederick S.; LIEBERMAN, Gerald J. **Introduction to operations research**. 11. ed. New York: McGraw-Hill Education, 2021.

KLEINROCK, Leonard. **Queueing systems: volume I – theory**. New York: John Wiley & Sons, 1975.

LAW, Averill M. **Simulation modeling and analysis**. 5. ed. New York: McGraw-Hill Education, 2015.

MA, Jing; WEBB, Timothy; SCHWARTZ, Zvi. A blended model of restaurant deliveries, dine-in demand and capacity constraints. **International Journal of Hospitality Management**, v. 96, art. 102981, 2021. DOI: 10.1016/j.ijhm.2021.102981.

MIGUEL, Paulo Augusto Cauchick. Estudo de caso na engenharia de produção: estruturação e recomendações para sua condução. **Produção**, São Paulo, v. 17, n. 1, p. 216–229, 2007.

PIEPER, Rafael; TANCREDI, Thiago; DÁVALOS, Ricardo Villarroel. Teoria das filas no planejamento portuário: um estudo de caso com o Terminal de Contêineres de Salvador. In: CONGRESSO INTERNACIONAL DE DESEMPENHO PORTUÁRIO – CIDESPORT, 12., 2025, Florianópolis. **Anais do XII Congresso Internacional de Desempenho Portuário**, Florianópolis, 2025.

RUBINSTEIN, Reuven Y.; KROESE, Dirk P. **Simulation and the Monte Carlo method**. 3. ed. Hoboken: John Wiley & Sons, 2017.

SHORTLE, John F.; THOMPSON, James M.; GROSS, Donald; HARRIS, Carl M. **Fundamentals of queueing theory**. 5. ed. Hoboken: John Wiley & Sons, 2018.

SLACK, Nigel; BRANDON-JONES, Alistair; JOHNSTON, Robert. **Administração da produção**. 8. ed. São Paulo: Atlas, 2018.

TAHA, Hamdy A. **Operations research: an introduction**. 10. ed. Harlow: Pearson, 2017.

USLUER, Furkan Onur. Solving queue problems in a campus dining service with discrete event simulation. **Ege Academic Review**, v. 25, n. 2, p. 351–360, 2025. DOI: 10.21121/eab.20250207.

VIEIRA, Victor Lindig; DÁVALOS, Ricardo Villarroel. Utilização da Teoria das Filas e Simulação Discreta para avaliar filas num Posto de Combustível da Região da Grande Florianópolis. In: CONGRESSO BRASILEIRO DE ENGENHARIA DE PRODUÇÃO – CONBREPRO, 2025. **Anais do Congresso Brasileiro de Engenharia de Produção**, 2025.