UNIVERSIDADE FEDERAL DO CEARÁ
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO, ATUÁRIA E CONTABILIDADE
DEPARTAMENTO DE ADMINISTRAÇÃO
CURSO DE CIÊNCIAS ATUARIAIS

FRANCISCO DANIEL SOUZA FERNANDES

TEORIA DAS FILAS E DELIVERY ONLINE: PROPOSTA DE UM MODELO GENERALIZADO PARA SISTEMAS COM ENTRADAS NÃO HOMOGÊNEAS

FORTALEZA
2023

FRANCISCO DANIEL SOUZA FERNANDES

TEORIA DAS FILAS E DELIVERY ONLINE: PROPOSTA DE UM MODELO GENERALIZADO PARA SISTEMAS COM ENTRADAS NÃO HOMOGÊNEAS

Monografia apresentada ao Curso de Ciências Atuariais da Faculdade de Economia, Administração, Atuária e Contabilidade da Universidade Federal do Ceará, como requisito parcial para obtenção do grau de Bacharel em Ciências Atuariais.

Orientador: Prof. Dr. Daniel Tomaz de Sousa

FORTALEZA

2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

F399t Fernandes, Francisco Daniel Souza.
TEORIA DAS FILAS E DELIVERY ONLINE : PROPOSTA DE UM MODELO GENERALIZADO PARA SISTEMAS COM ENTRADAS NÃO HOMOGÊNEAS / Francisco Daniel Souza Fernandes. – 2023.
34 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Faculdade de Economia, Administração, Atuária e Contabilidade, Curso de Ciências Atuariais, Fortaleza, 2023.
Orientação: Prof. Dr. Daniel Tomaz de Sousa.

1. Teoria das Filas. 2. Processo de Poisson não homogêneo. 3. variável no tempo. 4. delivery online. I. Título.

CDD 368.01

FRANCISCO DANIEL SOUZA FERNANDES

TEORIA DAS FILAS E DELIVERY ONLINE: PROPOSTA DE UM MODELO GENERALIZADO PARA SISTEMAS COM ENTRADAS NÃO HOMOGÊNEAS

Monografia apresentada ao Curso de Ciências Atuariais da Faculdade de Economia, Administração, Atuária e Contabilidade da Universidade Federal do Ceará, como requisito parcial para obtenção do grau de Bacharel em Ciências Atuariais.

Aprovada em: 08/12/2023.

BANCA EXAMINADORA

_____

Prof. Dr. Daniel Tomaz de Sousa (Orientador)
Universidade Federal do Ceará (UFC)

_____

Prof. Dra. Alane Siqueira Rocha
Universidade Federal do Ceará (UFC)

_____

Prof. Ma. Alana Katielli Nogueira Azevedo
Universidade Federal do Ceará (UFC)

A Deus, que sempre me amparou nos momentos mais difíceis, ainda que nem sempre eu tenha merecido...

A minha mãe e ao meu filho, por serem os motivos para eu continuar seguindo adiante.

AGRADECIMENTOS

À minha família, por ter me dado todo o apoio para que eu pudesse chegar até aqui, em especial minha mãe, Vanderléia, e ao meu irmão, Gabriel, que assistiu a defesa enquanto estava no trabalho.

Ao Prof. Dr. Daniel Tomaz de Sousa, meu xará, por toda a paciência e confiança com as entregas e sua atenção na orientação.

Às professoras participantes da banca examinadora, Prof. Dra. Alane Siqueira Rocha e Prof. Ma. Alana Katielli Nogueira Azevedo pelo tempo disponibilizado, pelas valiosas colaborações e sugestões para a versão final.

Aos amigos Alisson, Daniel e Marina, que ao longo dessa jornada me ajudaram imensamente, em praticamente todas as disciplinas, e que foram minha salvação em algumas das mais difíceis.

Ao Thiago, que ao me ver em necessidade de ajuda pesquisou sobre o tema e me deu uma dica essencial para conclusão do trabalho.

À Marieta, que realizou uma minuciosa revisão textual, assim como assistiu a primeira versão da apresentação e deu dicas de como melhorá-la para a “Hora H”.

RESUMO

A Teoria das Filas é uma ferramenta poderosa no auxílio à gestão de qualidade do serviço, contudo muitos problemas reais possuem diferentes particularidades as quais soluções algébricas não atendem completamente, como o caso das filas de entradas não homogêneas. Este estudo propõe uma combinação de dois métodos: um que estima a intensidade de chegadas no sistema quando esta segue um Processo de Poisson não homogêneo e outro que permite calcular as probabilidades de sistemas de fila com taxas de chegada variáveis no tempo, múltiplos servidores e limite de usuários simultâneos. A metodologia é desenvolvida a partir do estudo dos desafios de uma plataforma de *delivery* online, que intermedia os atendimentos de uma diversidade de estabelecimentos e consequentemente obtém-se um modelo generalizado, capaz de permitir a análise de sistemas diversos.

Palavras-chave: teoria das filas; Processo de Poisson não homogêneo; variável no tempo; delivery online

ABSTRACT

Queueing Theory is a powerful tool in aiding the management of service quality, however many real problems have different particularities which algebraic solutions do not fully meet, such as the case of non-homogeneous arrivals queues. This study proposes a combination of two methods: one that estimates the intensity of arrivals in the system when it follows a non-homogeneous Poisson Process and another that allows calculating the probabilities of queue systems with time dependent arrival rates, multiple servers and a limited number of simultaneous users. The methodology is developed based on the study of the challenges faced by an online delivery platform, which mediates services of a variety of businesses and therefore we obtain a generalized model, enabling analysis of various systems.

Keywords: queueing theory; non-homogeneous Poisson Process; time dependent; online delivery.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Otimização de custos é um objetivo comum a todas as organizações, bem como um tema de intrínseca complexidade, continuamente estudado por diversas organizações, na academia e fora dela. Ao longo do tempo, surgiram diversos campos cuja pretensão é munir os negócios de boas métricas e estratégias de gerir seus problemas foram construídas, seja na administração, probabilidade ou matemática aplicada. Nos últimos anos, o avanço tecnológico tem elevado a variedade e complexidade desses problemas, em rapidez e escala sem precedentes, o que multiplica os desafios enfrentados pelos estudiosos do tema.

Em várias áreas, como é o caso dos restaurantes, mercados, e lojas diversas, ganha cada vez mais espaço a operação por *delivery*, que acrescenta uma camada de complexidade operacional e logística, e as novas tecnologias tem tornado esse serviço imprescindível para a sobrevivência dos negócios. Nos últimos anos houve um boom no uso de aplicativos de *delivery*: apenas as operações da empresa iFood movimentaram R$ 31,8 bilhões em 2020, equivalente a 0,43% do PIB nacional (News iFood, 2021).

Segundo estimativas, no Brasil em 2022 havia mais de 250 aplicativos de entrega, e em 2021 este mercado movimentou aproximadamente R$ 35 bilhões no Brasil, o que representa 20% das vendas do setor de bares e restaurantes (Massa, 2022). É um mercado que envolve três partes: estabelecimentos, aplicativo e os entregadores. Em relação a estes últimos, o cenário econômico reflete a pressão a qual se submetem para garantir o meio de subsistência: 33% dos entregadores por aplicativo não tinham emprego quando iniciaram suas atividades na plataforma (Callil; Monise, 2023).

O cenário enfrentado pelas empresas no ramo atualmente é bastante complexo. Segundo estimativas, 93,4% dos negócios de restaurantes e lanchonetes no Brasil são micro e pequenas empresas (Magalhães, 2020), e em geral são empresas que enfrentam desafios operacionais diários, com ampla concorrência. Um dos principais novos componentes dessa interação entre cliente e loja é a plataforma online de pedidos, que age como agente intermediário e institui uma lógica própria que escapa ao controle dos estabelecimentos, ainda que remova deles a necessidade de alocação de entregadores. Uma descrição simples, mas precisa dessa estrutura foi ilustrada por Li *et al.* (2022).

A operação de minimizar os custos e simultaneamente manter os índices de qualidade mostra-se, sob essas circunstâncias, um desafio que demanda elevada cautela, posto que a própria exposição do negócio dentro da aplicação pode ser reduzida caso haja um aumento no

número de reclamações. Partindo disso, uma ferramenta adequada para a avaliação de um sistema de *delivery* pode ser realizada através da Teoria das Filas, uma vez que esta possui métodos próprios de mensuração de qualidade de serviço e abundante literatura em diversas áreas de aplicação.

Há diversos estudos que consideram simulações e análises em Teoria das Filas para a solução de problemas de casos específicos, mas que necessitam de certas adequações ad-hoc ao modelo. Gigante, Vieira e Azevedo (2021) aplicam um sistema para simular um processo de reabertura de comércio durante a pandemia de COVID-19. Lima *et al.* (2016) aplicam um modelo a um sistema de serviço de postagem. Em uma concepção comum em logística, Novaes (198) propõe modelos baseados em teoria das filas para a solução de problemas relacionados a dimensionamento de plataformas de carta e descarga.

Também são comuns aplicações da Teoria das Filas em soluções de entrega de comida. Um exemplo pode ser visto em Rahal e Yousef (2010), que reduzem o período de análise para ajustar o modelo como um Processo de Nascimento-e-Morte no período de maior movimento de uma pizzaria. Embora esse tipo de intervenção possa apresentar bons resultados, uma solução mais generalizada pode apresentar contribuições mais diversas. Zhang *et al.* (2019) propõem um sistema de taxa de entradas seguindo uma distribuição Poisson homogênea em uma rede de *fast-food*. Essa contribuição desconsidera casos em que o estabelecimento lida com horários de alto e baixo movimento, que na verdade são comuns.

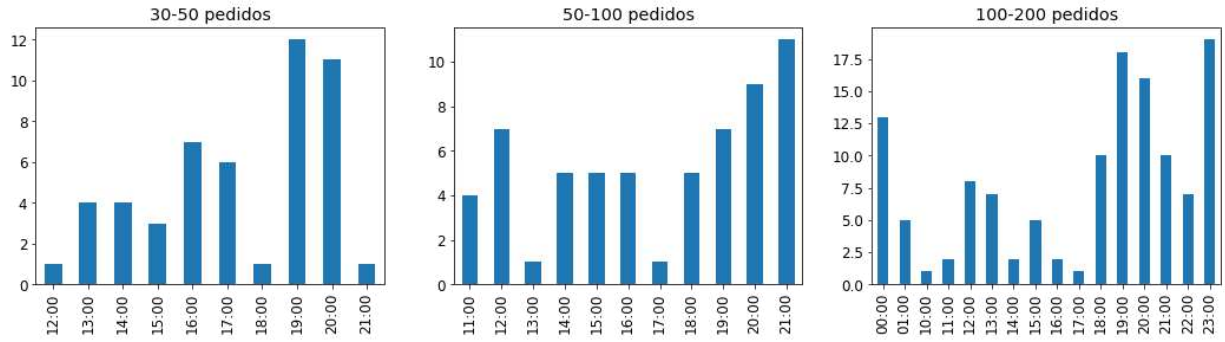
Dado o exposto, o objetivo geral deste trabalho é estudar um sistema de filas que contemple as necessidades reais de um estabelecimento com serviços de *delivery*, cujas propriedades possam ser aplicadas sem a necessidade de uma análise enviesada por suas limitações habituais, que forneça indícios adequados sobre a performance do estabelecimento em relação a seus indicadores de tempo de entrega e otimização dos serviços. Para tal, os objetivos específicos desta pesquisa são:

- Analisar um processo de estimação para a função $\alpha(t)$, a qual determina a taxa média de chegadas no sistema de filas de um Processo de Poisson Não-Homogêneo.
- Estudar um método de estimação para os tempos de serviço, dado que os dados analisados possuem limitações quanto à precisão do momento de início e término de cada entrega.
- Utilizar os dados obtidos e estimados pelos processos anteriores, a fim de simular um processo de filas $M_t/G/C/K$ aplicado a dados de *delivery* de uma plataforma online, o qual deverá fornecer informações de performance sobre o sistema de entregas de cada estabelecimento.

Como mostrado anteriormente, a vasta maioria dos estabelecimentos que utilizam serviços de *delivery* online é formada de pequenas e médias empresas, o que, no geral, indica que há recursos limitados de coleta e armazenamento de dados. Haja vista essas limitações, soluções de otimização de seus sistemas devem considerá-las, o que viabiliza a aplicação da Teoria das Filas como método analítico de performance. Este trabalho então justifica-se a partir dessa premissa.

A Figura 1 ilustra algumas das dificuldades de generalização do uso da Teoria das Filas no contexto de entregas de *delivery*, para uma amostra de três restaurantes em um dia específico: diferentes tipos de negócio têm diferentes horários de alto e baixo movimento, também os horários de funcionamento podem variar. Na maioria das vezes as distribuições dos processos de chegada seguem de forma não-homogênea.

Figura 1 – Amostra de chegada de pedidos para entrega



Fonte: Elaboração própria.

Soluções para sistemas de filas com taxas de entrada não homogêneos já foram propostas por diversos autores. Hasofer (1964) propôs um modelo para filas $M/M/1$ com entradas Poisson e tempo de serviço generalizado, Collings e Stoneman (1976) apresentaram uma proposta generalizada para Sistemas de Filas $M/M/\infty$ em tempo contínuo. Para propósitos de viabilidade computacional, será considerada aqui a abordagem de Brahimi e Worthington (1991), que traz uma aproximação para o tempo de serviço discreto para filas $M_t/G/C/K$ com capacidade finita, que pode ser considerado como um processo de *quasi* nascimento e morte. As premissas básicas deste modelo satisfazem as necessidades deste estudo.

No **capítulo 2**, será apresentada uma revisão da literatura de Teoria das Filas, bem como das premissas básicas do modelo que será seguido no restante do trabalho e suas formulações principais e cálculos pertinentes a esta aplicação. No **capítulo 3** serão discutidas algumas das propriedades do conjunto de dados disponibilizado para este trabalho, bem como suas peculiaridades. No **capítulo 4**, será feita uma avaliação dos processos de estimação dos

parâmetros discutidos no capítulo 2 quando aplicados aos dados apresentados no capítulo 3. No **capítulo 5** serão analisados pontos relacionados à questões técnicas do sistema de filas proposto, assim como será verificado se ele se comporta de modo adequado em comparação com os resultados de um outro trabalho e também ocorrerá uma discussão acerca de possibilidades de análises de performance para modelos de fila não-homogêneas. No **capítulo 6** serão realizadas considerações acerca dos resultados obtidos, com algumas reflexões sobre quais os possíveis próximos passos.

2 REFERENCIAL TEÓRICO

2.1 Um breve resumo de Teoria das Filas

As filas estão presentes em diversas situações da vida cotidiana: bancos, hospitais, centrais telefônicas, festas e restaurantes. Uma descrição típica do problema analisado em Teoria das Filas é dada por Albuquerque, Fortes e Finamore (2008, p. 273)

> Uma situação em que vários usuários de um serviço demandam este serviço de modo aleatório; o serviço é prestado por 1 ou mais servidores, sendo também em geral aleatório o tempo requerido para a prestação dele. Quando um usuário que chega ao sistema encontra todos os servidores ocupados, ele pode em geral esperar em uma fila até que um dos servidores desocupe.

Existem vários tipos de sistemas de filas, geralmente diferenciados através da representação proposta por Kendall, que caracteriza a distribuição de probabilidade do tempo de chegadas, a distribuição da duração do serviço, o número de servidores prestando atendimento, o número máximo de usuários permitidos no sistema (em atendimento ou em fila) e a população total de usuários (Albuquerque; Fortes; Finamore, 2008).

Um sistema de filas é frequentemente caracterizado por processos estocásticos da chegada e saída dos usuários neste, e suas relações mais úteis em aplicações práticas se encontram no Teorema de Little, o qual apresenta métricas de performance que independem das distribuições utilizadas no modelo, assumindo que o sistema ocorra na disciplina FCFS – *first-come-first-serve* (Albuquerque; Fortes; Finamore, 2008).

O comportamento de sistemas de filas pode ser expressado na forma das equações de Chapman-Kolmogorov, cujas derivadas são, segundo Albuquerque, Fortes e Finamore (2008), dadas por:

$$\frac{dp_0}{dt} = -\lambda_0 p_0(t) + \mu_1 p_1(t) \tag{1}$$

$$\frac{dp_k(t)}{dt} = \lambda_{k-1} p_{k-1}(t) - (\lambda_k + \mu_k) p_k(t) + \mu_{k+1} p_{k+1}(t) \tag{2}$$

Onde λ e μ são, respectivamente, as taxas médias de entrada e saída de usuários no sistema.

Estas equações descrevem as probabilidades para cada estado da cadeia de Markov em um determinado momento t. Embora haja, a princípio, diversas formas para solucioná-las, na

prática isso se torna muito dificultoso, conforme assumido por Worthington e Wall (1999) e Brahimi e Worthington (1991).

O sistema estudado a seguir assume a forma $M_t/G/c/K$ cuja sigla representa, respectivamente: com M_t, o sistema presume que as entradas de clientes na fila são dependentes do tempo, com G, o tempo de saída segue uma distribuição generalizada, C significa que o sistema possui capacidade para mais de um servidor em atendimento e L representa que há um limite máximo de usuários simultâneos no sistema.

Seja λ a taxa média de entrada de usuários em um sistema e β de o tempo médio de atendimento, alguns cálculos sobre o desempenho de um sistema de filas são: a duração média que os clientes permanecem no sistema $A = \lambda\beta$; o nível de serviço $S = \lambda\beta/m$; o tempo médio de retardo $E[d] = E[w] + E[y]$; onde w representa o tempo em espera e y o tempo em atendimento.

2.1.1 O sistema de filas de entradas não homogêneas proposto

O sistema de filas descrito neste trabalho especifica que o tempo t é dividido em m intervalos discretos e igualmente espaçados, e os tempos de serviço preenchem intervalos completos $1, 2, 3, \dots, m$. Esta metodologia é descrita em Galligher e Wheeler (1958), com aplicações em trabalhos subsequentes, como em Chassioti e Worthington (2004) e denominada DMT, ou Modelagem com Tempo Discreto. Aqui, um intervalo consistirá no espaço de um número de n minutos, e um evento cuja duração consuma uma parte de um intervalo contará como utilizando todo ele. Uma discussão extensa sobre a evolução literária da metodologia DMT e suas diferentes aplicações pode ser encontrada em Worthington e Wall (1999).

O sistema proposto por Brahimi e Worthington (1991) pressupõe a divisão do tempo de fila em $t = [0, T]$ períodos iguais e não sobrepostos. A cada intervalo t_j, com $j = 0,1,2,3, \dots, T$, o sistema é observado. A distribuição de chegada de pedidos em qualquer intervalo é independente dos outros, e os tempos de serviço devem ser identicamente distribuídos. O sistema possui C servidores, suporta no máximo L clientes simultâneos, com $L \geq C$, e em caso de ocupação total, o pedido entra em uma fila FCFS.

Adaptando a notação dos autores, os estados do sistema em um momento t são representados pelo conjunto de vetores $(n(t): X_1(t), X_2(t), \dots, X_m(t))$, no qual $n(t)$

representa o número de indivíduos no sistema no momento t, $X_i(t)$ representa o número X de indivíduos em serviço no momento t que permanecerão por mais i momentos, com $t \geq 0$. Por consequência, esse processo pode ser definido como uma cadeia de Markov com espaço de estados $j = k_1, k_2, \dots, k_m$, com $0 \leq j \leq L$ e $\sum_{i=1}^{m} k_i = \min(j, C)$.

Por conta da alta complexidade de solução das equações de Chapman-Kolmogorov, um algoritmo é desenvolvido considerando as possibilidades de estados para os vetores $(n(t): X_1(t), X_2(t), \dots, X_m(t))$, a partir desses vetores seguem-se as seguintes etapas:

1. são removidos os indivíduos que serão atendidos entre t e $t + \Delta t$, criando assim os vetores $(n_1(t): Y_1(t), Y_2(t), \dots, Y_m(t))$,
2. são adicionados $r = 0, 1, 2 \dots$ clientes a cada vetor $n_1(t)$, para todas as possibilidades de chegadas de novos indivíduos ao serviço, cujo número é denominado $nnewst$ e este é composto pelos vetores $(Z_1, Z_2, \dots, Z_m)$.
3. Ao incluir r clientes no sistema e iniciar o atendimento de $nnewst$ clientes, se obtém os vetores $(n_2(t): W_1(t), W_2(t), \dots W_m(t))$, cuja probabilidade é composta pelo somatório das probabilidades $P(X_1, X_2, \dots, X_m) * V_r(t) * P(Z_1, Z_2, \dots, Z_m)$, onde $V_r(t)$ denota a probabilidade de entrarem r clientes durante o momento t.

Esse algoritmo está descrito de forma detalhada em Brahimi e Worthington (1991) e aqui será desenvolvida uma adaptação dele escrita na linguagem Python.

2.2 A distribuição de entradas de pedidos

Devido às complexidades levantadas anteriormente, as chegadas de pedidos no sistema serão modeladas por um Processo de Poisson não-homogêneo, cuja definição é dada por Crestana (1991). O processo de Poisson é dado como $Y = \{Y(t); t \in [0, \infty]\}$, onde $Y(0) = 0$, Y tem incrementos independentes para todo t e sua função de probabilidade é dada por

$$P(Y(t) - Y(0) = x) = \frac{\left[\int_0^t \alpha(u)du\right]^x}{x!} e^{-\int_0^t \alpha(u)du} \qquad (3)$$

Onde $\alpha(t)$ é a função da taxa de ocorrência de um evento, conhecida como função intensidade. Se a função intensidade é integrável e toma a forma da função risco $\alpha(t) = f(t)/\big(1 - F(t)\big)$, então Watson e Leadbetter (1964 *apud* Ramlau-Hansen, 1983) introduzem um kernel estimador de sua intensidade, dado por

$$\hat{\alpha}(t) = \frac{1}{b}\sum_{i=1}^{n} \frac{K\left(\frac{t-X_i}{b}\right)}{n-i+1} \quad (4)$$

Este estimador pode ser generalizado na forma

$$\hat{a}(t) = \int_0^{\infty} \frac{1}{b} K\left(\frac{t-s}{b}\right) d\hat{\beta}(s) \quad (5)$$

Onde $\hat{\beta}(t) = \int_0^t \frac{1}{Y(s)} dN(s)$ é o estimador para a função de intensidade acumulada $\beta = \int_0^t \alpha(s)ds$.

Por questão de viabilidade computacional, para a estimação da função intensidade, seguir-se-á a equação sugerida por Ramlau-Hansen (1983), que assume a forma

$$\hat{\alpha}(t) = \frac{1}{b}\sum_{i=1}^{n} \frac{K\left(\frac{t-T_i}{b}\right)}{Y(T_i)} \quad (6)$$

Onde $T_i,\ i = 1,2,3,\ldots$ são os momentos dos eventos no processo empírico de contagem N, e $Y(t) = n - N(t-)$, ou seja, o total de eventos que ainda ocorrerão nos momentos $t, \ldots, T_n$.

Para a função K, será utilizado um kernel gaussiano, e o parâmetro b será selecionado através do método de validação cruzada da log-verossimilhança parcial, que pode ser escrita na forma

$$arg\max_b \ell(b,x) = -n\,log\,b + \sum_{i=1}^{n} log \sum_{i\neq j} K\left(\frac{T_i - T_j}{b}\right) \cdot \frac{1}{Y(T_i)} \quad (7)$$

2.3 Estimação dos tempos de serviço

Na metodologia adotada para o sistema proposto, não há pressuposição para o método de estimação da distribuição de probabilidade para os tempos de serviço, entretanto uma das limitações do conjunto de dados obtido é de que o horário do encerramento de uma entrega não é descrito de forma precisa, pois fica a cargo do estabelecimento o momento de registrar no sistema. Um dos propósitos deste trabalho é possibilitar a cada estabelecimento analisar sua performance mesmo sem acesso a dados históricos de longa duração, ou dados de outros estabelecimentos, portanto aqui também será considerado um método determinístico descrito por Novaes (1989), dado por

$$TC = t + T_p^* + T_\tau^* + t \quad (8)$$

Onde TC é o tempo de ciclo de uma entrega, que compreende o tempo total de uma rota, que pode conter uma ou mais entregas, T_p^* a soma dos tempos de paradas entre cada entrega em uma mesma rota, T_τ^* a soma dos tempos de deslocamento entre cada entrega na rota, e $t = 0$, e o tempo entre dois pontos é dado por $t_0 = d_0/v_0$ (Novaes, 1989).

Como o cálculo da rota ótima não faz parte do escopo desta pesquisa, assume-se de que cada rota consiste em apenas uma entrega, assim como $E[T_p^*]$ e $E[v_0]$ são constantes. Para o cálculo da distância d_0 será utilizada a Lei Esférica dos Cossenos.

Não seria irrazoável supor que outro estudo que utilize a metodologia aqui aplicada lide com dados de tempo de serviço mais precisos, portanto a aplicação de um método probabilístico também se faz bem-vinda. Assim sendo, outro método descrito será a utilização da estimação dos tempos de serviço através da distribuição gama, cuja função de densidade de probabilidade é dada por

$$\frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}} \qquad (9)$$

Sendo $\Gamma(k)$ uma função gama, dada por $\int_0^\infty t^{k-1}e^{-t}dt$, e os parâmetros k e θ são, respectivamente, a forma e escala da distribuição (Casella; Berger, 2010).

Os parâmetros k e θ serão estimados para cada estabelecimento com base na maximização da função de log-verossimilhança, utilizando as entregas intermediadas pelo próprio aplicativo e que não possuem imprecisão, realizadas para os estabelecimentos em um raio de até 2km de distância de cada estabelecimento amostral.

3 CONJUNTOS DE DADOS DO ESTUDO

Neste estudo será utilizado um conjunto de dados contendo os pedidos feitos através de uma plataforma online de *delivery*, realizados no dia 10 de janeiro de 2023 na cidade de Fortaleza, Ceará. Para avaliação de testes de hipóteses, serão considerados apenas estabelecimentos que atenderam a pelo menos 10 pedidos.

É importante salientar que os resultados deste estudo são obtidos em uma terça-feira, dia da semana que, historicamente, não costuma ser considerado como o mais movimentado para sistemas de *delivery*. Por haver uma diversidade em estabelecimentos e diferentes volumes de tráfego no conjunto de dados, entretanto, não se pressupõe, a princípio, que o método aqui desenvolvido não possa ser generalizado para outros dias da semana.

Por uma questão técnica no processo do próprio aplicativo, os estabelecimentos que possuem entrega própria não têm obrigatoriedade de registrar o campo Término Entrega de forma precisa, portanto, este campo não pode ser utilizado diretamente na estimação dos tempos de serviço. Ele pode, no entanto, ser utilizado na avaliação do nosso processo de estimação, ao verificar se a família de distribuições obtidas apresenta similaridades com aquelas dos estabelecimentos cuja entrega é feita pelo próprio aplicativo.

As variáveis disponíveis constam na Tabela 1, junto com um exemplo de preenchimento.

Tabela 1 – Variáveis do conjunto de dados

Ident. Pedido	**Estabelecimento**	**Tipo Estab.**	**Início Entrega**	**Término Entrega**	**Lat Cliente**	**Lon Cliente**	**Lat Estab.**	**Lon Estab.**
123	A	Pizza	2023-01-10 13:38:47.081738	2023-01-10 13:56:26.189754	-3.729	-38.585	-3.730477	-38.545755

Fonte: Elaboração própria.

Também será utilizado arquivo de dados espaciais no formato *Shapefile* contendo as coordenadas da cidade de Fortaleza, obtido no site Mapas de Fortaleza[1], mantido pela Prefeitura de Fortaleza (Ceará), para auxílio nas análises espaciais. Este arquivo também estará disponível no repositório do *GitHub*[2]. O *Shapefile* será utilizado para a avaliação da estimação dos tempos de entrega, bem como o limite L de clientes simultâneos no sistema.

[1] https://mapas.fortaleza.ce.gov.br/#/
[2] https://github.com/fdanielsouza/Teoria_das_Filas_Aplicado_a_*Delivery*_Online

4 ESTIMAÇÃO DOS PARÂMETROS DO MODELO

4.1 Avaliação do processo de estimação da função de intensidade

Seja α a intensidade de transição de estados de uma cadeira de Markov em nível individual, formulada sua estimação em (6), é sugerido por Ramlau-Ransen (1991) que a partir daí a decomposição do processo de contagem N pode ser definido como $A + M$, sendo A a integração do processo Λ, ou seja

$$A(t) = \int_0^t \Lambda(s)ds \qquad (10)$$

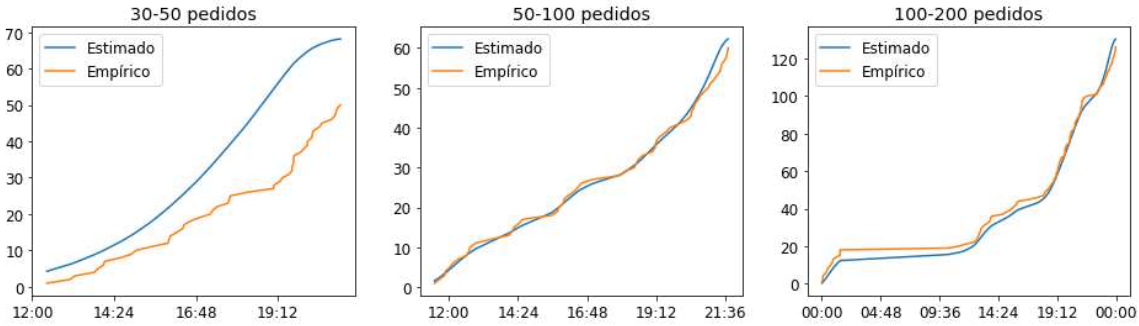
Em que Λ é definido como

$$\Lambda(t) = a(t)Y(t) \qquad (11)$$

Sendo M uma martingale que pode ser entendida como ruído no processo.

A análise gráfica do comparativo entre o processo A, gerado a partir das intensidades estimadas, contra o processo empírico N, sugere uma aderência considerável aos dados observados nos restaurantes com número de pedidos inferior a 30 e superior a 100 pedidos, mas uma superestimação do processo no restaurante com 30-50 pedidos.

Figura 2 – Comparativo do resultado de estimação X a distribuição empírica na amostra



Fonte: Elaboração própria.

De fato, realizado o teste de Kolmogorov-Smirnov para duas amostras nos 3 estabelecimentos observados, os respectivos p-valores são: 0,0; 0,9284 e 0,8244. Adotando um nível de significância de 0,05, pode-se rejeitar a hipótese nula no estabelecimento com 30-50 pedidos e, mas não se pode rejeitá-la nos outros. Quando generalizados os testes de Kolmogorov-Smirnov para todos os estabelecimentos, no entanto, não há sugestão de que o

problema esteja relacionado ao número de pedidos, mas a ocorrência de mudanças bruscas no tráfego de pedidos realizados aparenta influenciar este efeito, cabendo, porém, que algum outro trabalho possa se aprofundar neste ponto para identificar as causas.

Tabela 2 – Resultados do teste KS para os resultados estimados X os empíricos

Nº Pedidos	**10-30**	**30-50**	**50-100**	**100-200**	**> 200**	**Total**
% H_0 não rejeitado	96,79	67,95	42,86	55,56	0,00	85,33

Fonte: Elaboração própria.

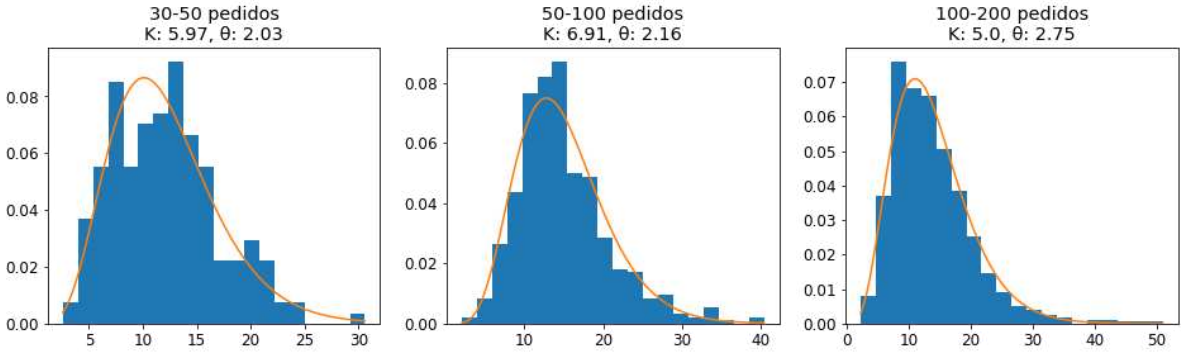
O alto percentual de testes de Komolgorov-Smirnov cujas hipóteses nulas não podem ser rejeitadas sugere que o estimador pode ser usado no processo aqui descrito com alto nível de confiança para estabelecimentos com 10-30 pedidos, porém casos individuais ensejam cuidados específicos, como uma mudança no método de escolha do parâmetro b, ou uma função Kernel diferente.

4.2 Aplicação do processo de estimação dos tempos de serviço

Outro requisito para um sistema de filas é a simulação de tempos de serviço, que pode seguir diferentes critérios. Como descrito na seção 2.3, aqui serão avaliados dois métodos de estimação para tempos de serviço, um determinístico, dado pela equação (8), e um probabilístico, dado em (9), que se refere à distribuição gama com parâmetros k e θ estimados através de estabelecimentos próximos ao que será avaliado.

Segundo a avaliação gráfica contida na Figura 3, há considerável aderência da família de distribuições teóricas às distribuições empíricas observadas, com os parâmetros em valores estimados pelo método de maximização da log-verossimilhança.

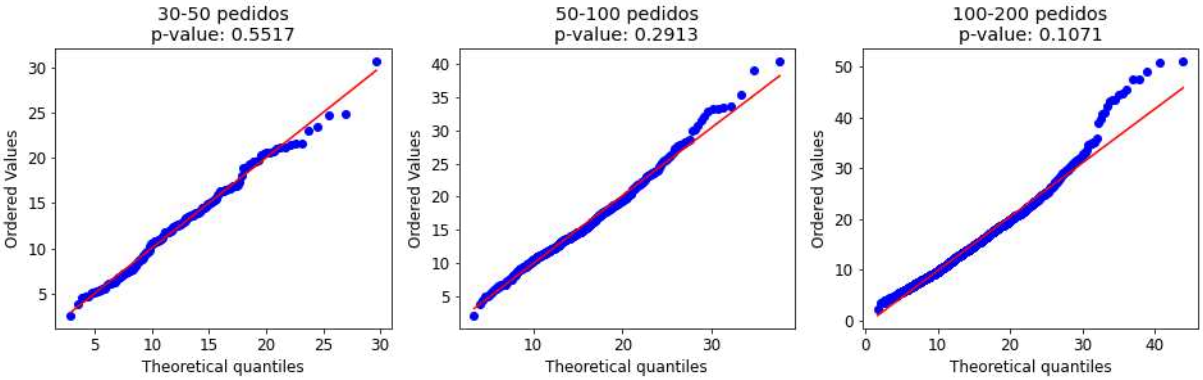
Figura 3 – Histograma de tempos de serviço X densidade da distribuição gama estimada



Fonte: Elaboração própria.

Conduzindo também um teste de aderência Kolmogorov-Smirnov, obtém-se os respectivos p-valores 0,5517; 0,2913 e 0,1071. Adotando um nível de significância de 0,05, não se pode rejeitar a hipótese nula para todos os estabelecimentos, assim podendo-se admitir que a distribuição é, de fato, adequada para a estimação desses dados. Avaliando o gráfico P-P mostrado na Figura 4, é possível perceber que, com exceção na cauda de tempos de serviço muito longos, o modelo de fato se adequa de forma satisfatória aos dados.

Figura 4 – Gráficos P-P dos pontos observados X a distribuição gama



Fonte: Elaboração própria.

Na Tabela 3, é possível observar o percentual de estabelecimentos os quais o p-valor do teste de Kolmogorov-Smirnov está acima de 0,05, com uma taxa de resultados acima de 80% na maior parte das segmentações de volume de pedidos, exceto para os estabelecimentos com mais de 200 pedidos.

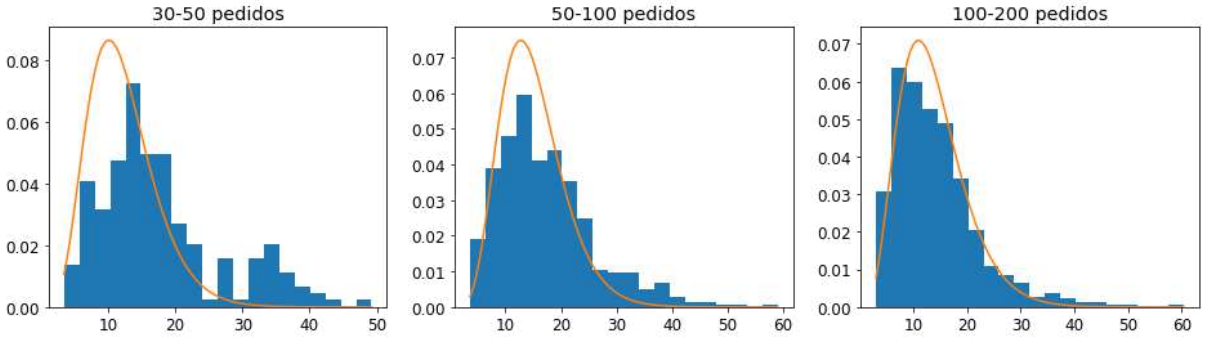
Tabela 3 – Resultados do teste KS para as estimações dos tempos de serviço

Nº Pedidos	10-30	30-50	50-100	100-200	> 200	Total
% H_0 não rejeitado	84,62	80,00	83,08	83,33	50,00	84,02

Fonte: elaboração própria.

Para a análise do método determinístico, caso seja necessário aplicar o modelo sem a obtenção de dados precisos de tempos de serviço, são necessários os parâmetros v_0 e T_p^*, que representam respectivamente a velocidade média em cada entrega e um tempo estimado de parada. Aqui esse parâmetro será ajustado visualmente, sendo o parâmetro de comparação para obtenção do valor ótimo o ajuste da curva da distribuição gama com os parâmetros calculados na subseção 4.2, e o melhor resultado obtido foi $v_0 = 10$ e $T_p^* = 3$, conforme exemplificado na Figura 5.

Figura 5 – Histograma de tempos de serviço estimados deterministicamente X curva da distribuição gama



Fonte: Elaboração própria.

Este método não apresentou resultados de estimação superiores aos da estimação a partir dos estabelecimentos dentro do raio de 2km, no entanto pode ser uma aproximação relativamente adequada dadas as possíveis limitações na coleta dos dados. A utilização deste método requer cuidados específicos de validação, tais quais análises gráficas, análises da particularidade dos percursos mais comuns das entregas, etc., a fim de que seus resultados não apresentem grande desvio do esperado por estimações mais tradicionais.

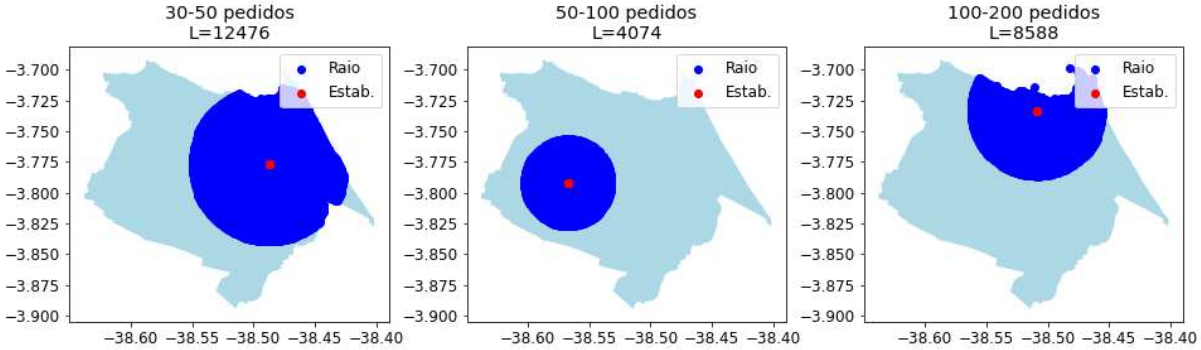
4.3 O número máximo de clientes permitidos no sistema

Tal qual apresentado na subseção 2.1.1, para que seja possível a implementação

computacional, o sistema de filas descrito por Brahimi e Worthington (1991) necessita também um parâmetro L, que determina um número máximo de clientes simultâneos no sistema, seja em espera ou em atendimento.

O número indicado para garantir plenamente o requisito deste parâmetro seria, naturalmente, o conjunto de todos os pontos de coordenadas geográficas que estão dentro do raio de entrega de determinado estabelecimento, truncadas em 3 casas decimais, conforme ilustrado na Figura 6.

Figura 6 – Raios de atendimento e valores para o parâmetro L



Fonte: Elaboração própria

Valores de L na casa dos milhares, entretanto, podem degradar bastante a performance do algoritmo, o que por sua vez pode impedir seu uso. Brahimi e Worthington (1991) sugerem que os requerimentos de processamento e armazenamento de dados podem aumentar rapidamente e de forma não linear, a depender dos valores utilizados para os parâmetros *C, m* e *L*, indicando também valores fixos para L que variam entre 30 e 80, a depender da intensidade de tráfego.

O parâmetro m, que representa o número máximo de unidades de tempo de serviço que um usuário pode estar em atendimento, de fato demonstrou alta sensibilidade para a degradação de performance. Na maioria dos casos, ao utilizar um Δt de 1 minuto resulta em algo próximo de $25 \leq m \leq 45$, que combinados com um valor alto de L e C podem elevar substancialmente o tempo necessário para rodar o programa.

Para efeito de garantia de performance computacional, o menor valor para L que não comprometa os cálculos de performance poderia ser estimado através de algum método, que foge ao escopo deste trabalho, atendendo ao requisito $L <= L*$, onde $L*$ representa esse número máximo no espectro de possibilidades.

5 APLICAÇÃO DO SISTEMA DE FILAS

5.1 Fluxo de funcionamento do algoritmo

Como resultado do algoritmo, obtém-se as probabilidades $P_t[n: x_m x_{m-1} \dots x_1]$, com $0 \leq n \leq L$ e $1 \leq x_i \leq C$, para todo $0 \leq t \leq T$ para um estabelecimento com sistema de filas qualquer. O fluxo seguido para a entrada dos parâmetros necessários se dá da seguinte forma:

1. Escolhem-se os pontos os quais o sistema será verificado. Ex.: se $\Delta t = 5\ min$, então são pontos analisados os horários 00:00:00, [...], 13:00:00, 13:05:00, 13:10:00, [...], 23:55:00.
2. Estima-se o parâmetro b da função de intensidade através do logaritmo de máxima verossimilhança, dada em (7).
3. Calcula-se o parâmetro de taxa de entradas λ através da fórmula (10) para todo t.
4. Os parâmetros ótimos para L e m são os que garantem que todas as possibilidades de se haver um número n de clientes aguardando pela duração máxima x_m. Estes números podem ser baseados na expectativa de um número máximo de usuários entrando no sistema ao mesmo tempo, multiplicado pelo tempo máximo de espera da distribuição empírica, por exemplo.
5. Estimam-se os parâmetros para a função de probabilidade de tempo de serviço $S_t(i)$.
6. Definem-se os valores de C para todo $0 \leq t \leq T$ com base em uma possível escala de disponibilidade de funcionários, ou lógica equivalente.
7. É executado o programa seguindo a lógica descrita por Brahimi e Worthington (1991) para cada t durante o período analisado.

Um programa em Python foi desenvolvido utilizando práticas de *trade-off* entre memória e CPU através da utilização de *Arrays* e *Iterators* para possibilitar a execução do algoritmo em um computador pessoal, e o código utilizado na construção do programa está disponível em repositório no *GitHub* criado para este projeto.

5.2 Validação do algoritmo

Para validação dos resultados do algoritmo apresentado na subseção 5.1, utilizaram-se os resultados encontrados no trabalho de Lima *et al.* (2016). Esse estudo avalia a qualidade do serviço de filas em uma agência dos correios ao longo de um período de 2 horas, utilizando um modelo de filas com 3 servidores, capacidade máxima de 41 clientes, com as chegadas seguindo uma distribuição Poisson homogênea e o tempo de espera uma distribuição Exponencial.

As métricas usadas para avaliar a precisão do algoritmo aqui apresentado consistem no número médio de clientes no sistema L_s, o número médio de clientes na fila L_q, o tempo médio em sistema W_s, as probabilidades dos estados $0, 1, \dots, 41$, utilizando-se as configurações $L = 41,\ m = 15,\ C = 3,\ k = 1$ e $\theta = 3.883$, sendo λ obtido através do processo descrito na subseção 4.1, e a comparação de resultados e seus desvios podem ser vistos na Tabela 4.

Tabela 4 – Comparativo dos resultados dos métodos

INDICADOR	LIMA *ET AL.*	RAMLAU-HAMSEN/BRAHIMI	DESVIO
MÉDIA DE CLIENTES NO SISTEMA	1,3336362	1,27345377	0,06018243
MÉDIA DE CLIENTES NA FILA	0,1039145	0,07357460	0,0303399
TEMPO MÉDIO NO SISTEMA	0,0350957	0,04659725	-0,01150155
PROB SISTEMA ESTAR VAZIO	0,2848143	0,30806505	-0,02325105
PROB SISTEMA NO ESTADO 1	0,3502428	0,30878323	0,04145876
PROB SISTEMA NO ESTADO 2	0,2153575	0,25835916	-0,04300919
PROB SISTEMA NO ESTADO 3	0,0882741	0,07523962	0,01303438

Fonte: Elaboração própria.

Os valores apresentados pelo método proposto por aproximação de tempo discreto se aproximam dos encontrados no modelo contínuo, com variações que podem ser consideradas aceitáveis para casos de aplicação tais quais o artigo propõe.

5.3 Considerações técnicas para a escolha dos parâmetros L, C e m.

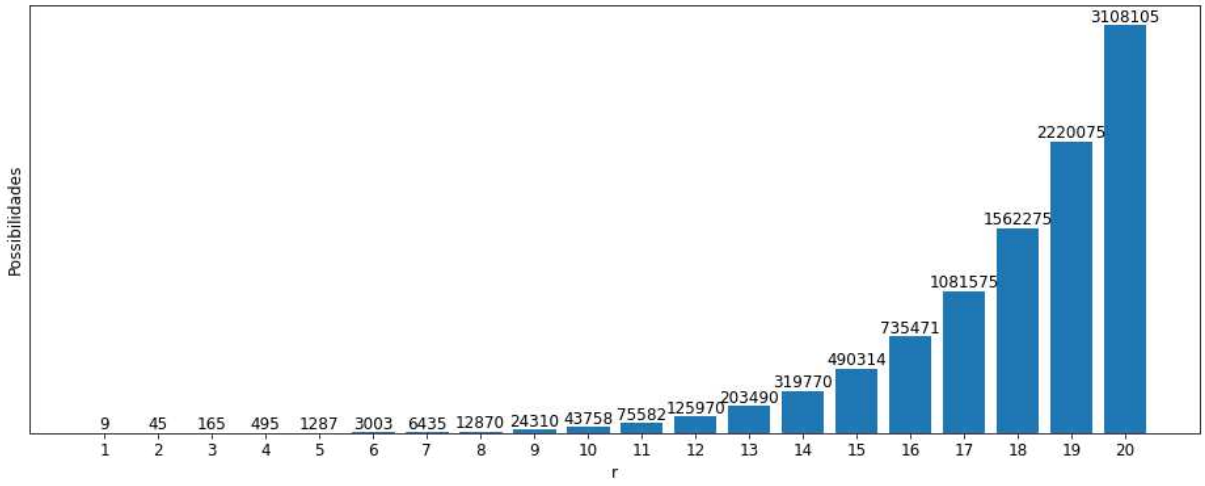
Ainda que se possa racionalizar um valor ideal para L de modo a cobrir todas as

possibilidades, como dado no subcapítulo 4.1, o desafio computacional persuade a utilizar números bastante inferiores, geralmente $L \leq 100$. Para a solução do problema de *delivery* analisada na subseção 5.4, o valor de m é escolhido através do número máximo de intervalos de 5 minutos de tempo de serviço necessário para um atendimento, utilizando todo o conjunto de dados amostral. Para o valor de L, considerar todas as possibilidades de duração de serviço para o número máximo de r entradas de novos pedidos ao longo do período analisado resultaria em uma aproximação mais precisa.

Para o conjunto de dados utilizado nos exemplos deste trabalho, decorre que $L = 63$, dado que $m = 9$. Brahimi e Worthington (1991) sugerem que para um alto volume de tráfego, $L = 80$ satisfaz as necessidades da maioria dos sistemas, enquanto $L = 30$ satisfaz um tráfego de baixo volume.

Para o parâmetro C, entretanto, um número muito alto pode inviabilizar uma rápida tomada de decisão, pois o número de combinações possíveis para os vetores $(n: x_m x_{m-1} \dots x_1)$ aumenta de forma exponencial, o que por sua vez gera um desafio computacional e deve ser feito um *trade-off* entre memória e tempo de espera. Através da Figura 7, é possível ver o número de possibilidades geradas para $1 \leq r \leq 20$:

Figura 7 – Número de vetores $X_1, X_2, \dots, X_m$ para cada chegada de r clientes no sistema



Fonte: Elaboração própria.

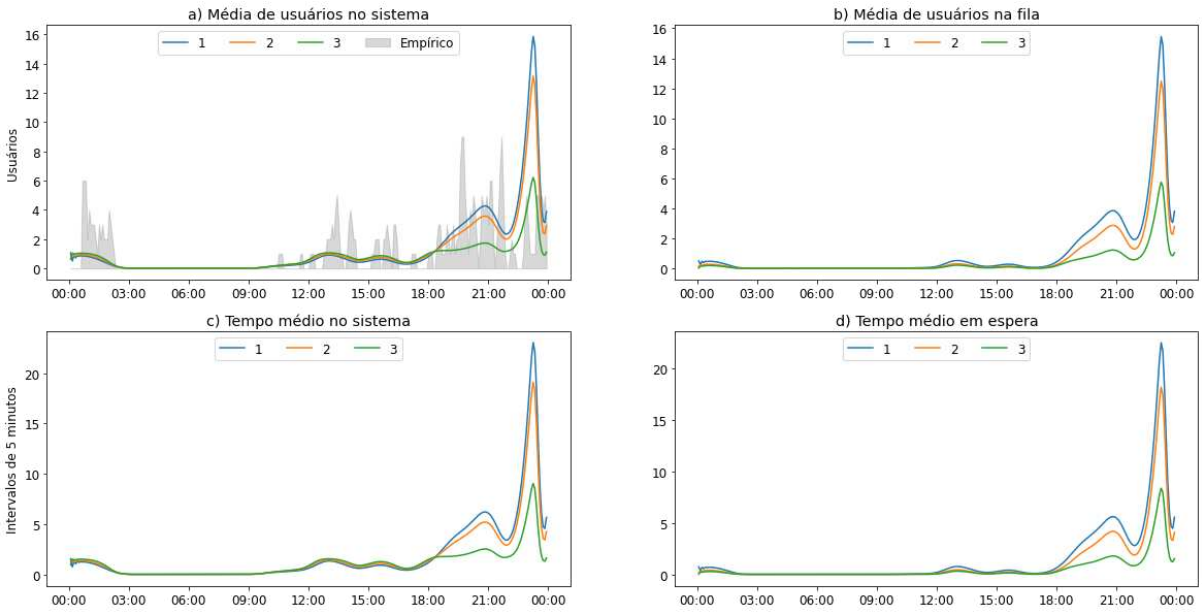
O algoritmo descrito, no entanto, de fato funciona eficientemente para números altos de L, C e m. O algoritmo descrito por Worthington e Wall (1999) precisa de 2 dias para rodar uma configuração com $L = 5$, $C = 5$ e $m = 20$, enquanto a mesma configuração aqui utilizada gera o espaço de estados $(n(t): X_1(t), X_2(t), \dots, X_m(t))$ em 453 milissegundos e executa todo o algoritmo para um t qualquer em menos de um minuto.

Computacionalmente, o algoritmo aqui utilizado, parametrizado através de uma configuração $L = 63$, $C = 5$ e $m = 9$ para qualquer t é viável em um computador equipado com um processador i7, com 3GHz de frequência e 16GB de memória RAM, mas precisa de cerca de 20 minutos em execução, consequentemente para concluir um dia inteiro nessas configurações são necessárias 96 horas em execução. É possível otimizar o uso do algoritmo através de determinadas particularidades dos estabelecimentos analisados, como por exemplo seu horário de funcionamento real.

5.4 Possibilidades para análises para o sistema de filas não-homogêneo

Embora neste modelo de fila não haja um ponto estacionário, a performance pode ser avaliada separadamente em cada t por meio das métricas descritas na subseção 2.1, podendo então serem identificados pontos de atenção e, consequentemente, uma ação mais focada. Na Figura 8 podem ser vistas as métricas de performance do sistema para $C = 1, 2, 3$, utilizando os dados do estabelecimento 100~200 pedidos:

Figura 8 – Métricas de performance de fila para $C = 1, 2, 3$



Fonte: Elaboração própria.

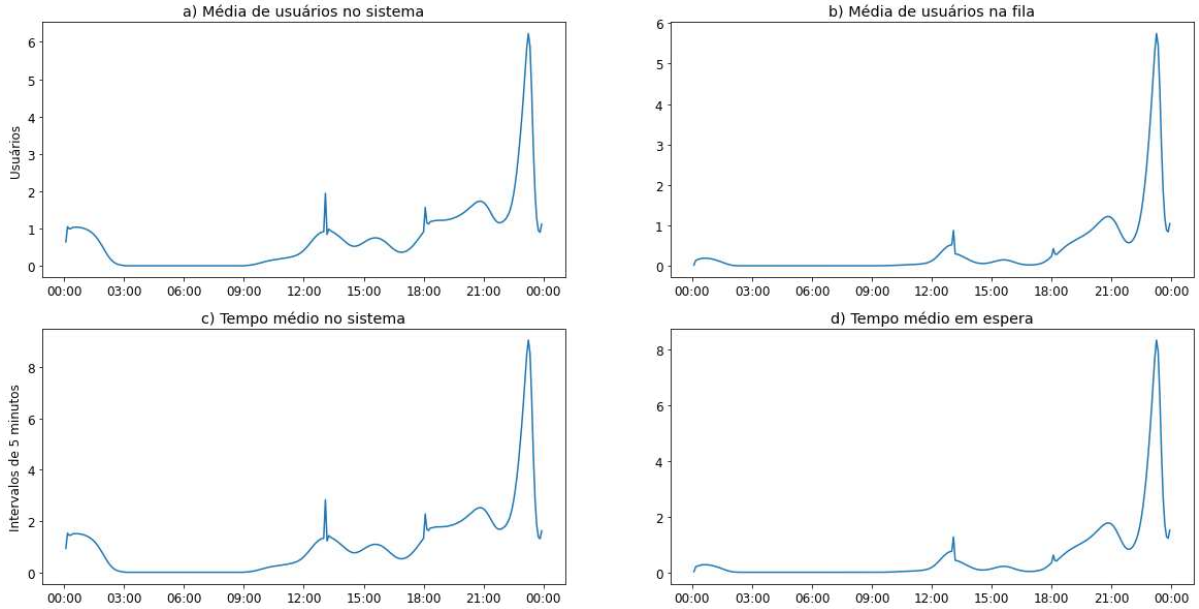
Através dos gráficos introduzidos na Figura 8, é possível perceber que manter $C = 1$

criaria um grande gargalo em usuários no sistema e em espera, conforme gráficos **a** e **b**. A utilização de 3 entregadores pode mitigar o problema em cerca de 80% entre os horários de 18 horas às 2 da manhã, mostrados no gráfico **b**. Uma análise dos tempos de espera no gráfico **d** indica que os tempos de espera podem chegar a 19 unidades de tempo de serviço quando $C = 2$, e mais que 27 quando $C = 1$.

Uma possível métrica global para avaliar a performance da fila poderia ser dada pelo percentual de unidades de tempo t nos quais há um número médio de clientes no sistema superior ao número de servidores ativos. No exemplo dado, para $C = 1$ temos 24,3% de unidades de tempo sobrecarregadas, para $C = 2$ e $C = 3$ esse percentual cai para 19,8% e 3,13%, respectivamente.

Como resultado das análises dos resultados gráficos, é compreensível que se proponha uma escala de entregadores, de tal modo a minimizar os impactos em performance e, concomitantemente, os custos. No exemplo utilizado, é possível propor, por exemplo: $C = 1$ entre 9 da manhã e 13 horas, $C = 2$ entre 13 e 18 horas e $C = 3$ entre 18 horas e 2 da manhã. A escala proposta incide em 3,13% de horários com capacidade média sobrecarregada, a mesma que mantendo $C = 3$ constante, e seus resultados gráficos são dados na Figura 9.

Figura 9 – Métricas de performance de fila com escala de funcionários



Fonte: Elaboração própria.

Embora não seja implementado no escopo deste trabalho, que é propor a combinação dos métodos de estimação para um processo de chegadas que segue uma distribuição Poisson não-homogênea e um sistema de filas que lide com essa particularidade, Wall e Worthington

(2007) propõem um algoritmo que visa a análise da performance de uma fila não homogênea através do tempo de espera virtual. Esse indicador funciona como métrica global e algébrica para determinar a qualidade da fila. Os autores consideram que esse algoritmo soluciona problemas encontrados nos métodos SSA (*Simple Stationary Approximation*), PSA (*Pointwise Stationary Approximation*) e MOL (*Modified Offered Load*).

6 CONSIDERAÇÕES FINAIS

Neste trabalho foi apresentado um método generalizado para filas no formato $M_t/G/c/K$, funcionando em conjunto com um método de estimação para a taxa de entradas, quando esta pode ser descrita como um Processo de Poisson não-homogêneo, através da estimação de sua função intensidade. Foi verificada a precisão do método através de comparação com os resultados obtidos em outro trabalho que utiliza o sistema de filas e apresentadas formas de avaliar a performance do sistema.

Originalmente, considerou-se aqui o problema dos sistemas de *delivery* online e suas particularidades, e o método descrito visa possibilitar a análise generalizada de diferentes estabelecimentos, ainda que funcionem de formas distintas. O principal problema visado pela solução aqui apresentada é a inconstância de atendimento em diferentes horários, assim como diferentes horários de funcionamento.

Além dos métodos supracitados, também foram discutidas alternativas para lidar com a falta de qualidade nas informações de tempos de serviço e foram apresentados métodos para a estimação destes, utilizando modelos estocásticos e determinísticos. Para examinar a aderência das estimações foi utilizado o teste de Kolmogorov-Smirnov, mas um estudo aprofundado acerca de testes alternativos poderia ser conduzido.

O método apresentado funciona para um modelo de fila simples, contudo Worthington e Wall (1999) discutem diversos outros sistemas de fila, com diferentes particularidades, assim como apresentam pesquisas que propõem a solução destas. Para lidar de modo adequado com esses diferentes desafios, o algoritmo aqui demonstrado pode atuar como uma base, mas poderão ser necessários ajustes com algoritmos citados na referida pesquisa.

Há também o desafio computacional. Ainda que a linguagem Python possa lidar de forma relativamente eficiente com os cálculos nas configurações citadas por Brahimi e Worthington (1991), conforme os valores dos parâmetros L, C e m aumentam, o tempo de execução aumenta de forma exponencial. Embora a solução proposta neste trabalho objetive possibilitar essa análise em computadores pessoais, uma possibilidade de pesquisa futura seria a adaptação do algoritmo para tecnologias mais escaláveis, como computação distribuída.

REFERÊNCIAS

ALBUQUERQUE, J. P. A.; FORTES, J. M. P.; FINAMORE, W. A. **Probabilidade, Variáveis Aleatórias e Processos Estocásticos**. Interciência - PUC-Rio, Rio de Janeiro, 2008.

BRAHIMI, M.; WORTHINGTON, D. J. **The finite capacity multi-server queue with inhomogeneous arrival rate and discrete service time distribution — and its application to continuous service time problems**. European Journal of Operational Research, v. 50, n. 3, p 310-324, fev, 1991. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/037722179190263U. Acessado em: 20 abr. 2023.

CALLIL, Victor; PICANÇO, Monise Fernandes. **Mobilidade urbana e logística de entregas: um panorama sobre o trabalho de motoristas e entregadores com aplicativos.** São Paulo: Centro Brasileiro de Análise de Planejamento Cebrap. 1. ed., 2023. Disponível em: https://cebrap.org.br/wp-content/uploads/2023/05/Amobitec12mai2023.pdf. Acesso em: 25 ago. 2023.

CASELLA, George; BERGER, Roger L. **Inferência Estatística.** Tradução da 2ª Edição Norte-Americana. Cengage Learning, São Paulo, 2010.

CHASSIOTI, E.; WORTHINGTON, D. J. **A New Model for Call Centre Queue Management**. The Journal of the Operational Research Society, v. 55, n. 12, p. 1352-1357, Dez, 2004. Disponível em: https://www.jstor.org/stable/4101854. Acessado em: 23 ago. 2023.

CRESTANA, Cassiana. **Um método de estimação para a função intensidade do processo Poisson não homogêneo**. Dissertação (Mestrado em Estatística) - Instituto de Matemática, Estatística e Ciência da Computação, UNICAMP. Campinas, 91 p. 1991. Disponível em: https://core.ac.uk/reader/296804220. Acessado em: 01 ago. 2023.

COLLINGS, T.; STONEMAN, C. **The M/M/∞ Queue with Varying Arrival and Departure Rates**. Operations Research, Glasgow, v. 24, n. 4, p. 760-773, Jul-Ago, 1976. Disponível em: https://www.jstor.org/stable/169773. Acesso em: 12 mai. 2023.

DIAS, Ronaldo. **Nonparametric Estimation: Smoothing and Visualization**. Departamento de Estatística, UNICAMP. Campinas, [s.d.]. Disponível em: https://www.ime.unicamp.br/~dias/SDV.pdf. Acessado em: 01 ago. 2023.

DRUCK, S. *et al*. **Análise Espacial de Dados Geográfico**s. EMBRAPA, Brasília, 2004.

FIPE revela que iFood responde por 32 bilhões do PIB nacional. **News iFood**, 2021. Disponível em: https://www.news.ifood.com.br/pesquisa-da-fipe-revela-que-ifood-responde-por-32-bi-do-pib-nacional. Acesso em: 20 jun. 2023.

GALLIGHER, Herbert P.; WHEELER, R. Clyde. **Nonstationary Queuing Probabilities for Landing Congestion of Aircraft**. Operations Research, v. 6, n. 2, p. 264-275, mar-abr, 1958. Disponível em: https://www.jstor.org/stable/167618. Acessado em: 18 ago. 2023.

GIGANTE, Rodrigo Luiz; VIEIRA, Henrique Ewbank de Miranda; AZEVEDO, Anibal

Tavarez de. **A relação entre o tempo de abertura do comércio, faturamento das lojas e exposição dos clientes ao COVID-19 com uso de Teoria das Filas**. Pesquisa Operacional para o Desenvolvimento, v. 14, p. 1-12, 2021. Disponível em: https://revistapodes.emnuvens.com.br/podesenvolvimento/article/view/685/449/ Acesso em: 26 ago. 2023.

HASOFER, A. M. **On the Single-Server Queue with Non-Homogeneous Poisson Input and General Service Time**. Journal of Applied Probability, v. 1, n. 2, p 369-384, dez., 1964. Disponível em: https://www.jstor.org/stable/3211866. Acessado em: 8 abr. 2023.

LI, J. *et al*. **Meal *delivery* routing optimization with order allocation strategy based on transfer stations for instant logistics services**. IET Intell. Transp. Syst. 16, p. 1108–1126, 2022. Disponível em: https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/itr2.12206. Acesso em: 22 jun. 2023.

LIMA, Katiucya Juliana Rodrigues de; MARTINS, Cassia Santos; DIAS, Glebton Daniel; COSTA, Luanna Barbosa. **Estudo da Teoria das Filas aplicado a uma empresa prestadora de serviços de postagem.** ENEGEP – XXXVI Encontro Nacional de Engenharia de Produção. Anais. João Pessoa, 2016. Disponível em: https://abepro.org.br/biblioteca/tn_stp_233_360_28859.pdf. Acesso em: 27 ago. 2023.

MAGALHÃES, Renata. Mercado de Restaurantes: Um Setor em Crescimento. **BuyCo**, 2020. Disponível em: https://buyco.com.br/mercado-de-restaurantes. Acesso em: 15 abr. 2023.

MASSA, Rubens Mussolin. **O "Boom" das plataformas de *Delivery* no Brasil e suas consequências peculiares**. Portal FGV, 2022. Disponível em: https://portal.fgv.br/artigos/boom-plataformas-*delivery*-brasil-e-suas-consequencias-peculiares. Acesso em: 28 ago. 2023.

NOVAES, Antônio Galvão. **Sistemas Logísticos: Transporte, Armazenagem e Distribuição Física de Produtos**. São Paulo: Ed. Edgard Blücher Ltda, 1989.

RAHAL, Ahmad D.; YOUSEF, Nabeel. **Service Optimization in the Fast Food Industry: The Case of the Pizza *Delivery* Service**. The Journal of Management and Engineering Integration, Orlando, v. 3, n. 1, p. 90-97, Summer, 2010. Disponível em: https://www.researchgate.net/profile/Ahmad-Rahal/publication/276914213_Service_Optimization_in_the_Fast_Food_Industry_The_Case_of_the_Pizza_*Delivery*_Service/links/555b42c808ae6aea0816994f/Service-Optimization-in-the-Fast-Food-Industry-The-Case-of-the-Pizza-*Delivery*-Service.pdf. Acesso em: 25 jul. 2023.

RAMLAU-HANSEN, Henrik. **Counting process intensities by means of Kernel functions**. The Annals of Statist, v. 11, n. 2, p. 453-466, jul. 1983. Disponível em: https://www.jstor.org/stable/2240560. Acesso em: 21 mar. 2023.

WALL, A. D. Worthington, D. J. **Time-dependent analysis of virtual waiting time behaviour in discrete time queues.** European Journal of Operational Research. v. 178, n. 2, p. 482-499, 2007. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S037722170600124X. Acesso em: 22 nov. 2023.

WORTHINGTON, Dave; WALL, Alan Douglas. **Using the discrete time modelling approach to evaluate the time-dependent behaviour of queueing systems**. The Journal of the Operational Research Society, v. 50, n. 8, p. 777-788, ago. 1999. Disponível em: https://www.jstor.org/stable/3010337. Acesso em: 18 set. 2023.

ZHANG, Tianhua; ZHAO, Fu; ZHANG, Juliang; MENDIS, Gamini; RU, Yihong; SUTHERLADN, John W. **An approximation of the Customer Waiting Time for Online Restaurants Owning *Delivery* System**. Journal of Systems Science and Complexity, v. 32, n. 32, p. 907-931, 2019. Disponível em: https://link.springer.com/article/10.1007/s11424-018-7316-4. Acesso em: 29 jul. 2023.