UNIVERSIDADE FEDERAL DE MINAS GERAIS
Escola de Engenharia
Programa de Mestrado em Engenharia de Produção

Alécio Rodrigues

APLICAÇÃO DE REDE DE FILAS MULTI-CLASSES PARA A REESTRUTURAÇÃO DE FLUXOS NOS SERVIÇOS HOSPITALARES DE URGÊNCIA E EMERGÊNCIA

Belo Horizonte
2025

Alécio Rodrigues

APLICAÇÃO DE REDE DE FILAS MULTI-CLASSES PARA A REESTRUTURAÇÃO DE FLUXOS NOS SERVIÇOS HOSPITALARES DE URGÊNCIA E EMERGÊNCIA

Dissertação apresentada ao programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Minas Gerais como requisito parcial para obtenção do título de Mestre em Engenharia de Produção.

Orientador: Prof. Dr. Maurício Cardoso de Souza

Coorientadora: Profa. Dra. Lásara Fabrícia Rodrigues

Belo Horizonte
2025

Rodrigues, Alécio.
R696a Aplicação de rede de filas multi-classes para a reestruturação de fluxos nos serviços hospitalares de urgência e emergência [recurso eletrônico] / Alécio Rodrigues. - 2025.
1 recurso online (69 f. : il., color.) : pdf.

Orientador: Maurício Cardoso de Souza.
Coorientadora: Lásara Fabrícia Rodrigues

Dissertação (mestrado) - Universidade Federal de Minas Gerais, Escola de Engenharia.

Inclui bibliografia.

1. Engenharia de produção - Teses. 2. Otimização - Teses. 3. Hospitais - Serviço de emergência - Teses. I. Souza, Maurício Cardoso de. II. Rodrigues, Lásara Fabrícia. III. Universidade Federal de Minas Gerais. Escola de Engenharia. IV. Título.

CDU: 658.5(043)

Ficha catalográfica pelo bibliotecário Reginaldo Cesar Vital dos Santos CRB6/2165
Biblioteca Prof. Mário Werneck, Escola de Engenharia da UFMG

UNIVERSIDADE FEDERAL DE MINAS GERAIS

Escola de Engenharia

Programa de Pós-Graduação em Engenharia de Produção

FOLHA DE APROVAÇÃO

Aplicação de Rede de Filas Multi-Classes para a Reestruturação de Fluxos nos Serviços Hospitalares de Urgência e Emergência

ALÉCIO RODRIGUES

Dissertação submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ENGENHARIA DE PRODUÇÃO, como requisito para obtenção do grau de Mestre em ENGENHARIA DE PRODUÇÃO, área de concentração PESQUISA OPERACIONAL E INTERVENÇÃO EM SISTEMAS SOCIOTÉCNICOS, linha de pesquisa Mod. e Algorit. de Otimiz. para Sistemas em Redes e de Prod..

Aprovada em 05 de fevereiro de 2025, pela banca constituída pelos membros:

Prof(a). Mauricio Cardoso de Souza - Orientador

UFMG

Prof(a). Lásara Fabrícia Rodrigues

UFMG

Prof(a). Marcus Vinicius Melo de Andrade

UFMG

Prof(a). João Flavio de Freitas Almeida

UFMG

Belo Horizonte, 05 de fevereiro de 2025.

Documento assinado eletronicamente por **Marcus Vinicius Melo de Andrade**, **Professor do Magistério Superior**, em 10/02/2025, às 17:49, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Mauricio Cardoso de Souza**, **Professor do Magistério Superior**, em 11/02/2025, às 15:11, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **João Flavio de Freitas Almeida**, **Professor do Magistério Superior**, em 11/02/2025, às 15:19, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Lásara Fabrícia Rodrigues**, **Professora do Magistério Superior**, em 11/02/2025, às 16:08, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3939930** e o código CRC **50F67E19**.

À Deus, aos meus queridos pais já falecidos, aos meus filhos, à minha esposa e aos meus irmãos.

AGRADECIMENTOS

A Deus, por nunca me abandonar, por estar sempre ao meu lado, fortalecendo minha fé e mostrando que, mesmo diante das dificuldades, tudo é possível.

Aos meus professores, Dr. Maurício Cardoso de Souza e Dra. Lásara Fabrícia Rodrigues, pela valiosa orientação, confiança depositada em mim, incentivo constante, paciência e amizade ao longo dessa jornada, além da generosa partilha de conhecimento e experiência.

À minha esposa, Daniela, e aos meus filhos, Thiago e Giovanna, pelo amor incondicional, compreensão e apoio em todos os momentos, o que me deu forças para seguir em frente.

Aos colegas da Excellence Health Solutions, pela confiança, parceria e compartilhamento de conhecimentos, que tornaram essa caminhada mais leve e significativa.

À minha funcionária Andreia, pelo convívio diário, incentivo contínuo e torcida incansável, que foram fundamentais neste processo.

Por fim, a todos que, de alguma forma, contribuíram direta ou indiretamente para a realização desta dissertação, deixo registrado o meu mais profundo e sincero agradecimento.

“Cada sistema está perfeitamente planejado para obter os resultados que obtém”. (Paul B. Batalden, 2015)

RESUMO

No Brasil tornou-se natural encontrar pacientes com dias de espera nos serviços de urgência, sobretudo públicos, configurando uma situação dramática, infelizmente naturalizada. Os serviços de urgência têm dificuldade de internar seus pacientes e com isto se transformam em grandes enfermarias. A segmentação dos pacientes, a partir da classificação de risco, separando os pacientes de baixo, médio e alto risco clinico e, a inclusão de três novas áreas para melhor organização dos fluxos (*fast-track,* teleconsulta e unidade de decisão clínica) são algumas das estratégias adotadas na redução dos tempos de espera.

Esta dissertação, uma pesquisa-ação, tem como objetivo promover o alinhamento entre a capacidade instalada dos recursos (locais de cuidado e profissionais de saúde) do pronto-socorro hospitalar, tratado neste caso como uma rede de filas aberta (*OQN*), e a demanda dos pacientes, observando a qualidade do serviço prestado, bem como a segurança do paciente. O modelo captura diferenças específicas do hospital no perfil de risco clínico dos seus pacientes, padrões e volumes de chegada e, eficiência dos processos em um único modelo computacional.

Uma implementação das equações de filas é utilizada, para dimensionar as áreas do pronto-socorro usando o tempo de espera, o fator de utilização e a probabilidade de superlotação como metas de qualidade de serviço. A aplicação da metodologia resultou em uma redução significativa nos tempos de espera, permitindo uma resposta mais rápida e eficiente às emergências médicas. A eliminação de desperdícios e a otimização no fluxo dos pacientes não apenas aumentaram a eficiência operacional, mas também melhoraram substancialmente a experiência do paciente. A equipe assistencial, agora, melhor estruturada, consegue oferecer um atendimento de qualidade superior.

Palavras-chave: redes de filas abertas; curvas de *trade-off*; otimização.

ABSTRACT

In Brazil, it has become common to find patients waiting for days in emergency services, especially public ones, representing a dramatic situation that has unfortunately been normalized. Emergency services struggle to admit patients, often turning into large inpatient wards. Strategies such as segmenting patients through risk classification—separating them into low, medium, and high clinical risk—and introducing three new areas for better flow management (fast-track, teleconsultation, and clinical decision units) are some of the strategies adopted to reduce waiting times.

This dissertation, an action research, aims to promote the alignment between the installed capacity of resources (care facilities and healthcare professionals) in the hospital emergency department, treated here as an Open Queueing Network (OQN), with patient demand while ensuring service quality and patient safety. The model captures hospital-specific differences in the clinical risk profiles of patients, arrival patterns and volumes, and process efficiency within a single computational framework.

Queueing equations are implemented to optimize the emergency department's areas using wait time, utilization factor, and overflow probability as service quality targets. Applying this methodology resulted in a significant reduction in wait times, enabling faster and more efficient responses to medical emergencies. Waste elimination and optimized patient flow not only enhanced operational efficiency but also substantially improved patient experience. The better-structured care team now delivers superior-quality service.

Keywords: open queue network; trade-off curve; optimization.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ALT Alta médica do pronto-socorro

CID Classificação Internacional de Doenças

CQN Rede de filas fechada (*Closed Queueing Network*)

CRG Centro cirúrgico (Intervenção cirúrgica de urgência)

D2D Tempo porta-médico (*Door to Doc*)

DES Simulação de eventos discretos (*Discrete-event simulation*)

DOC Médico (*Doctor*)

EDBA *Emergency Department Benchmarking Alliance* (www.edbenchmarking.com)

ERP Software de Planejamento de Recursos Empresariais (*Enterprise Resource Planning*)

ESI *Emergency Severity Index*

FCFS Primeiro a chegar, primeiro a ser servido (*First Come, First Served*)

FTK Fluxo rápido (*Fast-Track*)

INT Internação (enfermaria, unidade de tratamento intensivo)

LAB Laboratório (exames de laboratório)

LOH Tempo médio de internação (*Lenght Of Hospitalization*)

LOS Tempo médio de permanência (*Lenght Of Stay*)

LWBS Pacientes que deixam o PS sem serem vistos pelo médico (*Left Without Been Seen*)

MTS Protocolo de Manchester (*Manchester Triage System*)

NPS Índice de satisfação do cliente (*Net Promoter Score*)

OM Administração da produção (*Operations Management*)

OQN Rede de filas aberta (*Open Queueing Network*)

OR Pesquisa Operacional (*Operational Research*)

PS Pronto-Socorro

RAD Radiologia (exames de imagem)

REG Registro do paciente

SADT Serviço de Apoio Diagnóstico Terapêutico

SVM Sala Vermelha (*Shock Room*)

TELE Teleconsulta

TRG Classificação de risco (Triagem)

TRT Tratamento (medicação, sutura, gesso, entre outros)

UDC Unidade de decisão clínica

UPA Unidade de Pronto Atendimento

SUMÁRIO

1. INTRODUÇÃO

1.1. Contextualização

No Brasil tornou-se natural encontrar pacientes com dias de espera nos serviços de urgência, configurando uma situação dramática, infelizmente naturalizada. Os serviços de urgência têm dificuldade de internar seus pacientes e com isto se transformam em grandes enfermarias. Pesquisa de opinião, Figura 1, realizada no Brasil, aponta a saúde como preocupação constante e que em alguns momentos supera a violência, a corrupção, educação e o desemprego como um dos piores problemas do país na ótica da população.

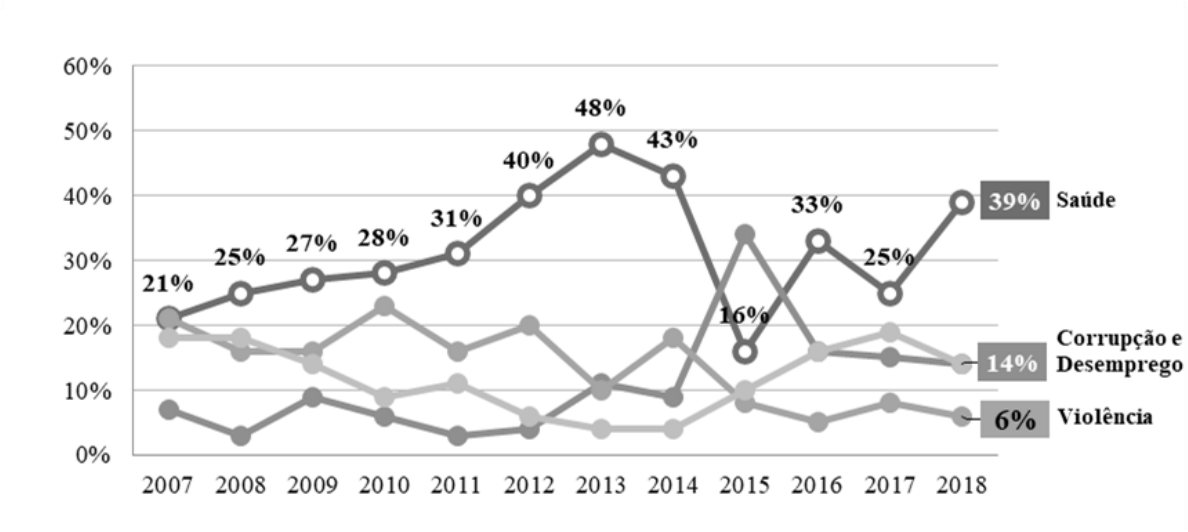


Figura 1 – Saúde como principal problema do país (fonte: CFM/Datafolha, 2018)

A pesquisa realizada pelo Instituto Datafolha (2018), a pedido do Conselho Federal de Medicina (CFM), demonstrou que a percepção negativa sobre a saúde no Brasil, tanto na esfera privada quanto pública continuava alta, entre os brasileiros. A área era vista como péssima e regular por 89% da população, e boa parte desta carga negativa decorria da experiência vivenciada por 150 milhões de pessoas que dependiam exclusivamente da rede do SUS para fazer uma consulta, um exame ou uma cirurgia. Essa insatisfação decorria, sobretudo, do longo tempo de espera para que o cidadão obtivesse uma resposta para sua demanda associado à falta de recursos financeiros para o SUS e a inadequada gestão administrativa e operacional, assim como a falta de médicos e a dificuldade para marcar ou agendar consultas, cirurgias e procedimentos, entre outras questões.

Na área da saúde, os prontos-socorros são a porta de entrada de pacientes em situação de urgência e emergência. Em relação a esse serviço, essa mesma pesquisa apontou que 63% dos entrevistados avaliaram o atendimento nos prontos-socorros e UPAs como regular, ruim ou péssimo (Datafolha, 2018).

No Brasil, a situação se agrava a passos largos. Pelo lado da demanda, existe um aumento crescente da procura pelos prontos-socorros impulsionada inicialmente pelas questões demográficas, haja visto que a expectativa até 2035, é de dobrar a taxa de idosos, alcançando 20% da população, feito este que a França levou 150 anos para alcançar, conforme Figura 2.

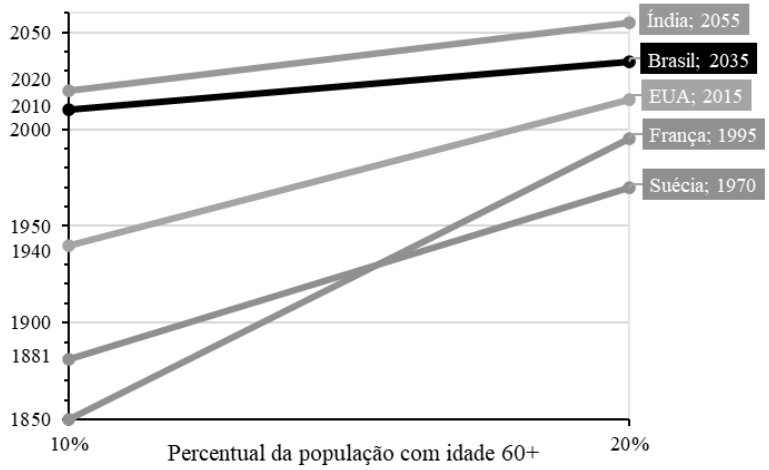


Figura 2 – Em 25 anos, Brasil dobrará a taxa de idosos, alcançando 20% da população
(fonte: OMS – Velocidade do envelhecimento populacional, 2020)

Outra questão a ser observada, conforme Figura 3, diz respeito ao aumento das doenças crônicas (câncer, diabetes, problemas cardiovasculares, entre outras) enquanto, as doenças infecciosas, como a Dengue, Zika e Chikungunya, ainda estão presentes e, além disso, as causas externas, como a violência interpessoal. Por último, mas não menos importante, o fato de os prontos-socorros serem vistos, sobretudo pelos pacientes de baixo risco clínico, como um local de conveniência, ou seja, solução para problemas simples, como o pedido de receitas ou exames e afecções de baixa complexidade.

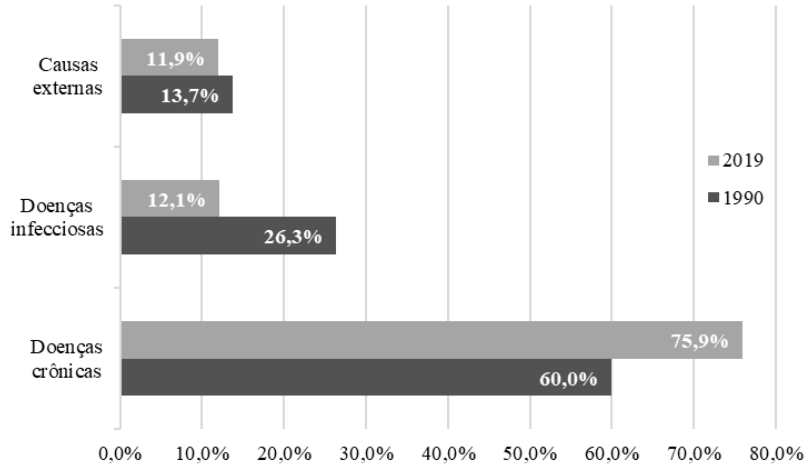


Figura 3 – Distribuição das principais causas de morte
(fonte: GBD Brasil, Estudo Carga Global de Doenças, 2019)

Já pelo lado da oferta, como ilustrado pelas Figuras 4 e 5, observou-se uma redução progressiva dos serviços entre 2009 e 2017 que alcançou 8,02% do número de leitos por 1 mil habitantes e 3,7% no número de hospitais. Portanto, o cenário é complexo e a solução deve ser a de melhorar a eficiência dos recursos disponíveis.

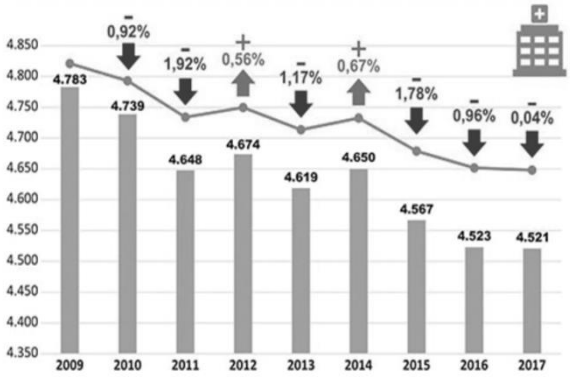


Figura 4 – Evolução do núm. hospitais públicos e privados disponíveis ao SUS (fonte: CNES)

Figura 5 – Evolução da rede hospitalar brasileira de 2009 a 2017 (fonte: CNES)

A superlotação dos prontos-socorros, definida como uma saturação do limite operacional dos serviços de urgência tem como sua causa principal a dificuldade de internar os pacientes (EVIPNet Brasil, 2020), provocando longas esperas (*boarding*). Uma revisão sistemática (Filippatos e Evridiki, 2015) mostrou que: (i) o tempo de *boarding* aumenta a mortalidade em até 2 vezes (a partir de 12 horas de permanência); (ii) o tempo de internação hospitalar aumenta de 4 a 7 dias numa correlação importante com o tempo de *boarding* no pronto-socorro (quanto mais horas no PS, mais dias internado no hospital) e; (iii) aumento dos eventos adversos (50% das causas de erros num hospital). Trata-se, portanto, de um desafio contínuo e crítico para a eficiência operacional no mundo inteiro e tem sido associada a efeitos negativos tanto para os pacientes quanto para os profissionais. Os pacientes que procuram atendimento em prontos-socorros lotados estão sujeitos a maiores riscos de morbidade e mortalidade (Derlet e Richards, 2002), tempo de espera prolongado (Derlet e Richards, 2000), maior probabilidade de deixar o serviço sem ser atendido (evasão/abandono) e taxas mais altas de insatisfação (Derlet e Richards, 2000 e 2002; Sprivulis *et al.*, 2006). Do ponto de vista do provedor, a superlotação pode levar a taxas mais altas de eventos adversos (Gordon *et al.*, 2001), falta de comunicação e estresse, bem como menor produtividade e moral. Além disso, a superlotação pode ter efeitos negativos na missão de ensino em hospitais universitários e reduzir a capacidade dos prontos-socorros de responder a incidentes com múltiplas vítimas.

Os hospitais respondem à superlotação do pronto-socorro utilizando basicamente dois tipos de intervenção para acomodar os picos de demanda: (i) segmentação do fluxo de pacientes pelo risco clínico (gravidade) e por especialidade, comumente chamado de "*fast-track*" e, (ii) utilização dos espaços de espera para cumprir parte da internação do paciente (*boarding*).

A segmentação do fluxo de pacientes visa tratar pacientes de baixo risco clínico em uma fila separada daqueles pacientes de maior risco que aguardam pela sua internação hospitalar em um leito, muitas vezes, improvisado na urgência. Separar os pacientes de baixo risco (menor gravidade) faz com que os mesmos não concorram pelos mesmos locais de cuidado (leitos, poltronas, macas, entre outros locais dedicados à assistência) do pronto-socorro. Ela atua como uma espécie de atalho para os pacientes, em torno do problema da espera pelo leito. Muitas variações dessa ideia já foram testadas e, geralmente, sob a bandeira do "*fast-track*".

Normalmente, os pacientes são acomodados para aguardar em uma "maca no corredor" ou de volta à recepção depois de se consultarem com um médico. Os desenhos mais arrojados fazem com que os pacientes de menor risco caminhem entre as áreas de tratamento onde são avaliados por um médico, recebem procedimentos e aguardam pelo resultado dos exames, se necessário. O benefício dessa ideia de projeto é que a maioria dos pacientes passa pouco tempo em um local de cuidado no pronto-socorro, aumentando assim a capacidade destes locais para tratamento dos pacientes. O objetivo é reduzir o tempo, em que um paciente de menor risco clínico, esteja consumindo recursos das áreas de atendimento médico.

A incorporação dessas intervenções melhora o desempenho do pronto-socorro tanto operacionalmente quanto em termos de segurança do paciente. Pesquisas anteriores (Goodacre e Webster, 2005 e Rowe, 2006) citam longos tempos porta-médico como fator principal para pacientes abandonarem o serviço. Para unificar e adaptar esses princípios a um determinado pronto-socorro hospitalar, resta o problema de como realizar o planejamento da capacidade para essa nova geração de fluxos dentro do pronto-socorro. O método deve ser geral, mas incorporar a combinação de risco clínico dos pacientes a serem atendidos, os padrões de volume de chegada destes pacientes e os períodos de tempo de consumo de recursos dentro de um pronto-socorro específico.

Desta forma, a segmentação pode ser utilizada para ajustar a capacidade de atendimento do pronto-socorro à demanda dos pacientes. A segmentação dos pacientes, a partir da classificação de risco, separando os pacientes de baixo, médio e alto risco clinico e, a inclusão de três novas áreas para melhor organização dos fluxos (*fast-track – FTK*, teleconsulta – *TELE* e unidade de decisão clínica – *UDC*) são algumas das estratégias adotadas. Valendo destacar

que, os recursos não foram aumentados e nem tampouco diminuídos. E, a partir da definição da qualidade requerida para prestação do serviço, em termos de tempo de espera, taxa de utilização dos recursos e probabilidade de superlotação (*overflow*), os recursos foram alocados, em cada uma das áreas, de forma a garantir o melhor atendimento à demanda.

Além disto, com base na Classificação Internacional de Doenças (*CID*), é possível identificar quais são as doenças ou queixas mais prevalentes do pronto-socorro. E, dada esta informação, construir, para aqueles pacientes potencialmente candidatos ao *fast-track*, protocolos clínicos bem definidos, atrelados a um menu de exames de resposta rápida, cujo tempo seja inferior a 30 minutos, devidamente pactuados com o SADT, de forma que a equipe médica possa tomar decisões rápidas e eficazes, baseadas em evidencias científicas e melhores práticas clínicas. Isso não apenas melhora a qualidade do atendimento prestado, mas também reduz a variabilidade no cuidado ao paciente, minimizando eventos adversos, aumentando a segurança e evitando exames que não definem a conduta. Portanto, trata-se, também esta, de uma estratégia fundamental para melhorar o atendimento médico e promover uma gestão mais eficiente dos recursos hospitalares.

1.2. Objetivos

1.2.1. Geral

Esta dissertação tem como objetivo promover o alinhamento entre a capacidade instalada dos recursos (locais de cuidado e profissionais de saúde) do pronto-socorro, tratado neste caso como uma rede de filas aberta (*OQN*), e a demanda dos pacientes, observando a qualidade do serviço prestado, bem como a segurança do paciente. O modelo captura diferenças específicas do hospital no perfil de risco clínico dos seus pacientes, padrões e volumes de chegada e eficiência dos processos em um único modelo computacional. Portanto, o objetivo deste trabalho é apresentar um novo paradigma de atendimento ao pronto-socorro que reduza as “evasões” e aumente o acesso ao pronto-socorro, por meio de um método de pesquisa operacional que se adapta a qualquer hospital através do uso de dados específicos e normalmente disponíveis no hospital.

1.2.2. Específicos

Dentre os objetivos específicos, destacam-se:

- Melhorar a experiência do paciente por meio da redução: da taxa de evasão (*LWBS*), do tempo porta-médico (*D2D*) e do tempo médio de permanência (*LOS ou leadtime*);

- Reduzir o tempo médico-decisão através do ajuste na demanda por exames laboratoriais e de imagem (*SADT*) promovida pela segmentação dos pacientes e o redesenho dos fluxos;
- Reduzir o tempo decisão-saída (*boarding*) com o ajuste da capacidade;
- Aumentar a receita através da redução: da taxa de evasão (*LWBS*), da demanda por exames e do tempo de *boarding*;
- Dimensionar as áreas (recursos) do pronto-socorro usando o tempo de espera, o fator de utilização e a probabilidade de superlotação (*overflow*) como metas de qualidade de serviço e;
- Explorar e evidenciar o potencial da análise de curvas de *trade-off* para o planejamento dos serviços de urgência.

1.3. Estrutura do trabalho

Esta dissertação está organizada conforme o esquema da Figura 6.

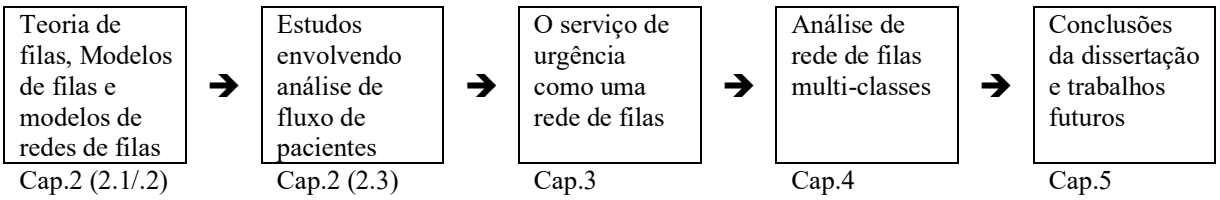


Figura 6 – Organização dos capítulos da dissertação

O capítulo 2 apresenta os conceitos sobre teoria de fila referenciados ao longo do texto, a seção 2.1 trata dos principais modelos de filas, a seção 2.2 dos modelos de redes de filas e a seção 2.3 dos estudos envolvendo análise de fluxo de pacientes.

O capítulo 3 discute sobre como um serviço de urgência pode ser representado por redes de filas abertas (*OQN*), apresenta o estado anterior à intervenção (*as-was*) da pesquisa-ação, principais indicadores operacionais e o fluxo de pacientes.

O capítulo 4 descreve como o serviço de urgência do hospital "H" pode ser melhorado. A seção 4.1 apresenta a nova configuração da rede de filas, na seção 4.2 descreve-se o método proposto e apresenta os resultados computacionais, na seção 4.3 são apresentadas as curvas de *trade-off* e suas respectivas análises e, na seção 4.4 são apresentadas algumas estratégias para melhorar o giro de leito nas unidades de internação e com isso mitigar o problema de *boarding* no pronto-socorro.

Finalmente, o capítulo 5 apresenta as conclusões desta dissertação e trabalhos futuros.

2. TEORIA DE FILAS

Os estudos sobre filas foram iniciados pelo engenheiro dinamarquês Agner K. Erlang que publicou seu primeiro trabalho em 1917 aplicando a teoria de probabilidade ao problema de tráfego de linhas telefônicas (Brockmeyer *et al.*, 1948). O objetivo inicial era estudar o congestionamento e os tempos de espera que ocorriam no momento em que as ligações telefônicas eram realizadas. Após este estudo, muitos outros têm sido desenvolvidos em diversas áreas como por exemplo, comunicação digital, hospitais, oficinas mecânicas, redes logísticas, dentre outras.

2.1. Modelos de filas

Um sistema de fila, de acordo com Morabito, R. (1998), é definido por um fluxo de clientes ou tarefas que chegam e são tratados por um servidor. Se o servidor estiver ocupado atendendo um cliente quando um segundo chega, este último deve aguardar pelo serviço. Após serem servidos, os clientes deixam o sistema.

A Figura 7 ilustra essa representação. Convém salientar que os modelos de filas são motivados por casos em que os processos de chegada ou de serviço, ou ambos, são probabilísticos, resultando possivelmente numa fila de espera de clientes. Portanto, cada nó (ou estação) contém os seguintes elementos: (i) processo de chegada, (ii) processo de serviço e (iii) fila de espera.

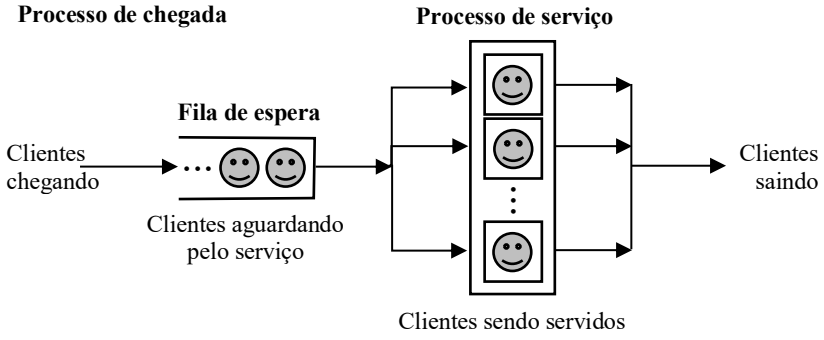


Figura 7 – Um sistema de fila de único estágio (fonte: adaptada de Morabito, 1998)

O **processo de chegada** na estação é descrito pelo intervalo de tempo entre chegadas de clientes, podendo ser determinístico ou probabilístico. Se o processo de chegada for probabilístico, ele pode ser dependente dos outros intervalos de tempo entre chegadas e/ou do processo de serviço, ou consistir de intervalos entre chegadas independentes e identicamente distribuídos (*iid*). O primeiro caso é chamado de processo de chegada *G* (genérico), já o

segundo caso, processo de chegada *GI* (genérico independente), é aquele em que o número de chegadas ao longo do tempo ocorre conforme um processo de *Poisson*, isto é, os intervalos de tempo entre chegadas são independentes e exponencialmente distribuídos (processo *Markoviano*, ou sem memória, M).

Os clientes podem pertencer a uma única classe ou família, neste caso, todos os clientes são supostos idênticos estatisticamente ou, agrupados em múltiplas classes diferentes. Em geral, clientes cujas demandas são parecidas, consomem recursos similares podem ser agrupados em classes. Há várias razões para a agregação de dados, talvez a mais óbvia seja reduzir a complexidade computacional dos modelos, que depende do número de classes e estações. Mas há também a preocupação de reduzir a complexidade gerencial: um gestor diante de inúmeras variáveis provavelmente terá dificuldades para identificar os processos dominantes que governam o comportamento do sistema.

O **processo de serviço** na estação é descrito pelo tempo de atendimento dos clientes, que pode ser determinístico ou probabilístico. Se o processo de serviço for probabilístico, ele pode depender de outros tempos de atendimento e/ou do processo de chegada, ou consistir de tempos de atendimento *iid*. Alguns autores têm chamado o primeiro caso de processo de serviço *G* e o segundo caso de processo de serviço *GI* (Disney e Konig, 1985). Aqui, ambos os casos são referidos apenas como processo de serviço *G*, uma vez que é mais usual na literatura. Um exemplo de processo de serviço *G* dependente do processo de chegada é aquele em que o tempo de atendimento varia de acordo com o número de clientes na fila. Similarmente aos intervalos de tempo entre chegadas, os tempos de atendimento podem ser variáveis aleatórias independentes e exponencialmente distribuídas (processo de serviço *Markoviano*, ou sem memória, M).

A **fila de espera** na estação pode ter sua capacidade limitada, geralmente determinada pelo espaço físico disponível e uma vez alcançada, a entrada de novos clientes na fila é bloqueada ou, ilimitada. A fila tem uma disciplina ou regra para ordenar como os clientes serão atendidos. Alguns exemplos de disciplinas são: primeiro a chegar, primeiro a ser servido (*first come, first served – FCFS*), último a chegar, primeiro a ser servido (*last come, first served – LCFS*), serviço de ordem randômica (*served in random order – SIRO*) e, fila com prioridades: primeiro o de menor tempo de atendimento (*shortest processing time first – SPT*) e, primeiro o de maior tempo de atendimento (*largest processing time first*).

No caso de fila com prioridades, pode-se ter o caso preemptivo e o não-preemptivo. No caso preemptivo, o cliente com maior prioridade entra em serviço assim que chegar na fila,

mesmo que um cliente com menor prioridade já esteja em serviço. No caso não-preemptivo, um cliente já iniciado não pode ser interrompido enquanto não for completado.

Na maioria dos modelos de filas supõe-se que as chegadas e as saídas do sistema ocorrem de acordo com o processo de nascimento e morte. Assim, no contexto da teoria das filas, o nascimento se refere às chegadas de clientes no sistema e a morte às partidas de clientes já servidos. Este processo de nascimento e morte é importante para descrever probabilisticamente como o estado do sistema muda de acordo com o tempo. De maneira geral, nesses processos, nascimentos e mortes individuais ocorrem aleatoriamente e suas taxas médias de ocorrência dependem somente do estado atual do sistema.

As suposições deste processo são:

(i) dado um número n de clientes no sistema no tempo t ($t > 0$) a distribuição de probabilidade atual do tempo restante até a chegada de outro cliente é exponencial com parâmetro λ_n ($n = 0, 1, 2, 3...$);

(ii) dado um número de clientes na fila no tempo t ($t > 0$) a distribuição de probabilidade atual do tempo restante até a conclusão do serviço seguinte é exponencial com parâmetro μ_n ($n = 1, 2, 3...$);

(iii) somente um nascimento ou morte pode ocorrer de cada vez. Portanto, ao considerar as suposições (i) e (ii) pode-se dizer que os modelos baseados no processo de nascimento e morte são aqueles nos quais o processo de chegada possui distribuição de *Poisson* e o processo de serviço exponencial.

Como para um processo de nascimento e morte pode-se atribuir qualquer valor não negativo às taxas médias de chegada e de serviço, existe uma grande flexibilidade na modelagem de um sistema de filas. Contudo, este processo oferece apenas um ajuste razoável para alguns tipos de filas. Reconhecendo a diversidade dos sistemas de filas, em Kleinrock (1976), pode-se caracterizá-los usando a notação proposta por David Kendall, em 1953, e descrita por uma série de símbolos, tais como, A/B/c/N/K/Z (ver por exemplo em Kleinrock, 1976) onde:

A: Distribuição do tempo entre chegadas

B: Distribuição do tempo de serviço

c: Número de servidores (atendentes paralelos)

N: Capacidade máxima do sistema (número máximo de clientes no sistema)

K: Tamanho da população

Z: Disciplina da fila

Também existe uma notação condensada, dada por apenas três termos (A/B/c), em que se supõe que não haja limite para o tamanho da fila, que a população seja infinita e a que disciplina da fila seja FIFO (*first in, first out*), ou seja, o primeiro a entrar é o primeiro a sair.

Modelo M/M/1

O modelo de fila M/M/1 é um dos modelos mais simples e amplamente utilizados em teoria das filas. Ele descreve o comportamento de um sistema de fila onde as chegadas de clientes ao sistema seguem um processo de *Poisson*, o que significa que as chegadas ocorrem de forma aleatória e independente, com uma taxa média de chegada denotada por λ, e representa quantos clientes, em média, chegam ao sistema por unidade de tempo. O tempo entre as chegadas segue uma distribuição exponencial.

O atendimento ou serviço aos clientes também segue um processo de *Poisson*, independentemente das chegadas, com uma taxa média de serviço denotada por μ, e representa quantos clientes, em média, o servidor é capaz de atender por unidade de tempo. O tempo necessário para atender a um cliente segue uma distribuição exponencial.

Este modelo pressupõe que há apenas um servidor disponível para atender os clientes na fila. Esse modelo, portanto, é um ponto de partida fundamental para a análise de sistemas de filas mais complexos e é amplamente utilizado em várias aplicações, como redes de computadores, *call centers*, sistemas de transporte, entre outros.

No modelo de fila M/M/1, várias medidas de desempenho podem ser calculadas para avaliar o comportamento do sistema. As principais medidas de desempenho são mostradas na Tabela 1:

Medida de desempenho	**Descrição**	**Equação**
Taxa de utilização do sistema (ρ)	é a fração de tempo que o servidor está ocupado atendendo clientes. É uma importante medida pois indica o quão eficientemente o servidor está sendo utilizado. Quanto maior a utilização, mais ocupado o servidor está;	$\rho = \frac{\lambda}{\mu}$
Tempo médio de espera na fila (W_q)	indica o tempo médio que um cliente passa na fila esperando para ser atendido pelo servidor;	$W_q = \frac{\lambda}{\mu.(\mu - \lambda)}, \rho < 1$

Tempo médio de espera no sistema (W)	representa o tempo médio que um cliente passa no sistema, incluindo o tempo na fila e o tempo de serviço;	$W = \frac{1}{(\mu - \lambda)}, \rho < 1$
Número médio de clientes na fila (L_q)	indica a média do número de clientes na fila esperando para serem atendidos pelo servidor;	$L_q = \frac{\lambda^2}{\mu.(\mu - \lambda)}, \rho < 1$
Número médio de clientes no sistema (L)	representa a média do número de clientes, incluindo aqueles na fila e sendo atendidos pelo servidor;	$L = \frac{\lambda}{(\mu - \lambda)}, \rho < 1$
Probabilidade de o sistema estar vazio (P_0)	é a probabilidade de não haver clientes no sistema, ou seja, a probabilidade de o servidor estar ocioso;	$P_0 = 1 - \rho, \rho < 1$
Probabilidade de n clientes no sistema (P_n)	representa a probabilidade de haver exatamente n clientes no sistema.	$P_n = (1 - \rho).\rho^n, \ n \geq 0$

Tabela 1 – Principais medidas de desempenho (M/M/1)

Essas medidas de desempenho são essenciais para avaliar o funcionamento de um sistema de fila M/M/1 e podem ser usadas para otimizar a capacidade do sistema, dimensionar servidores adicionais, reduzir o tempo de espera dos clientes e melhorar a eficiência operacional.

Modelo M/G/1

O modelo de fila M/G/1 é uma extensão do modelo M/M/1 que acomoda distribuições de tempo de serviço gerais, tornando-o adequado para modelar sistemas de filas com tempos de serviço não exponenciais, como aqueles encontrados em muitos cenários do mundo real, como atendimento ao cliente, assistência médica, e muitos outros.

As medidas de desempenho no modelo de fila M/G/1 são semelhantes às medidas usadas no modelo M/M/1, mas devido à natureza geral da distribuição de tempo de serviço (G), a análise é frequentemente mais complexa e, muitas vezes, requer técnicas numéricas ou simulações para serem calculadas com precisão.

Dado que os clientes chegam ao sistema de acordo com um processo de *Poisson*, de taxa λ e que o tempo de serviço (conhecidos o valor médio x e a variância σ_x^2) não necessariamente segue uma distribuição exponencial, não se pode afirmar que o processo aleatório que descreve

o número de clientes no sistema em um instante de tempo qualquer, $k(t)$, seja Markoviano.

Considerando $\rho = \lambda\bar{x}$ e, após algumas manipulações matemáticas, chega-se à fórmula de Pollaczed-Khinchin (P-K), para o cálculo do número médio de clientes no sistema E[k]. E, usando a fórmula de Little, $E[k] = \lambda E[T]$, é possível derivar as demais equações de desempenho, conforme mostrado na Tabela 2.

Medida de desempenho	**Descrição**	**Equação**
Número médio de clientes no sistema $E[k] = L$	representa a média do número de clientes, incluindo aqueles na fila e sendo atendidos pelo servidor;	$E[k] = \rho + \frac{\rho^2 + \lambda^2\sigma_x^2}{2(1-\rho)} = \rho + L_q = L$
Tempo médio de espera no sistema $E[T] = W$	representa o tempo médio que um cliente passa no sistema, incluindo o tempo na fila e o tempo de serviço;	$E[T] = \bar{x} + \frac{\rho^2 + \lambda^2\sigma_x^2}{2\lambda(1-\rho)} = \bar{x} + W_q = W$
Tempo médio de espera na fila $E[W] = W_q$	indica o tempo médio que um cliente passa na fila esperando para ser atendido pelo servidor;	$E[W] = \frac{\rho^2 + \lambda^2\sigma_x^2}{2\lambda(1-\rho)} = \frac{L_q}{\lambda} = W_q$
Número médio de clientes na fila $E[N_q] = L_q$	indica a média do número de clientes na fila esperando para serem atendidos pelo servidor;	$E[N_q] = \frac{\rho^2 + \lambda^2\sigma_x^2}{2(1-\rho)} = L_q$

Tabela 2 – Principais medidas de desempenho (M/G/1)

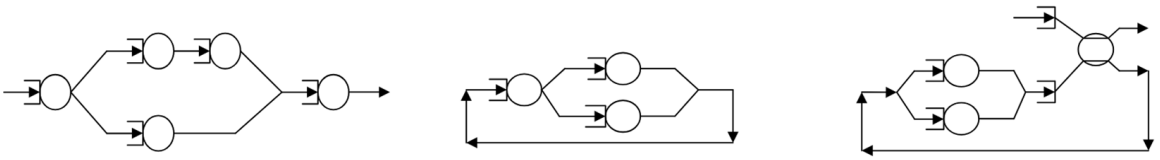
2.2. Modelos de redes de filas

Os modelos de redes de filas vêm sendo amplamente utilizados na análise do desempenho de sistemas complexos, tais como rede de computadores, sistemas de comunicação, de saúde, de manufatura, entre outros. Koenigsberg (1982), Disney e Konig (1985), Wein (1990b), Kouvelis e Tirupayi (1991), Calabrese (1992), Hsu *et al.* (1993), Suri *et al.* (1993), Gershwin (1994), Britan e Sarkar (1994a), Morabito (1998), Silva (2005), Silva e Morabito (2007) são alguns exemplos de estudos relativos às redes de filas.

Define-se uma rede de filas como um arranjo de centros de serviço, em que um conjunto de entidades denominadas clientes (podendo ser, pacientes) recebem atendimento (serviço) em uma ou mais estações. O centro de serviço consiste de um ou mais servidores, que

correspondem aos recursos do sistema modelado, e uma área de espera para que os clientes possam aguardar pelo serviço. É, portanto, necessário estudar toda a rede para poder extrair dela informações sobre número médio de clientes no sistema (L), tempo médio de permanência no sistema (W), número médio de clientes aguardando pelo serviço (L_q) e, tempo médio de espera, antes de ser atendido (W_q).

O conjunto de nós, arcos e clientes compõe a rede de filas com as seguintes características: (i) número de estações (nós), (ii) sequência de atendimentos (roteiros) e, (iii) tipologia: aberta, fechada e mista (Figura 8).



a) Rede de filas aberta b) Rede de filas fechada c) Rede de filas mista

Figura 8 – Tipologia de redes de filas

O número de nós na rede, corresponde ao número de estações. Cada estação pode ser visitada pelo mesmo cliente mais de uma vez, e em cada visita pode realizar um atendimento diferente. A sequência de visitas (ou roteiro) ao longo das estações, determinística ou probabilística, pode ser sequencial, sequencial com realimentação (*feedback*), arborescente, acíclica e cíclica.

Em uma rede de filas aberta (*open queueing networks – OQN*), os clientes assim que chegam entram na rede, recebem atendimento em um ou mais nós, e eventualmente saem da rede. A Figura 9 ilustra uma rede aberta em que clientes entram na rede pela estação "E" do lado esquerdo da Figura, percorrem roteiros sequenciais com realimentação (ou cíclicos), ao longo das estações (conforme indicado pelos arcos da Figura), eventualmente aguardam em fila defronte às estações visitadas, e deixam a rede pela estação "S" do lado direito da Figura. O número de clientes circulando na rede é uma variável aleatória.

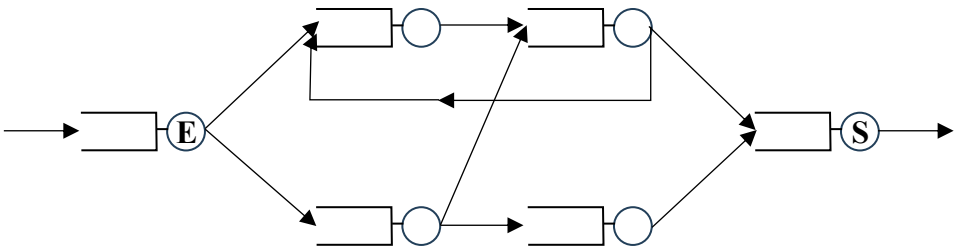


Figura 9 – Rede de filas aberta (*OQN*)

Em uma rede de filas fechada (*closed queueing network – CQN*), ao contrário, não há chegadas ou partidas externas, o número de clientes circulando na rede é mantido fixo, enquanto a taxa de partidas internas de cada nó é uma variável aleatória. Entretanto, pode-se artificialmente representar chegadas e partidas externas de clientes numa *CQN*, ao se definir uma estação de carga e descarga instantânea. Para isso, para cada partida externa de um cliente desta estação, há uma chegada externa de outro cliente nesta mesma estação e, portanto, o número de clientes na *CQN* é controlado para que se mantenha constante. Se a rede tiver múltiplas classes de clientes, pode-se redefinir sub-redes abertas para algumas classes e sub-redes fechadas para outras. Neste caso, a rede de filas é chamada mista.

O modelo de rede de filas analisado nesta dissertação refere-se a *OQN*, além de serem analiticamente mais tratáveis do que modelos *CQN*, podem ser utilizados para aproximá-los (veja Whitt, 1984 e Calabrese, 1992). Os modelos pressupõem que o sistema atinja o estado de equilíbrio ou regime permanente (*steady state*), isto é, baseiam-se no comportamento do padrão típico do sistema num longo período de tempo. Em outras palavras, os resultados dos modelos descrevem o comportamento médio de longo prazo de sistemas "estáticos", ou seja, sistemas em que seja razoável admitir que parâmetros como a taxa média de chegada de clientes ou o tempo médio de atendimento de clientes não variam ao longo do horizonte de análise (processos de chegada e de serviço estacionários). Também se admite que os sistemas sejam estáveis no sentido de que suas capacidades, medidas pela taxa máxima de produção, excedam a demanda média, de maneira que os níveis médios de WIP (clientes em processo) ou os tempos médios de espera sejam finitos.

2.3. Estudos envolvendo análise de fluxo de pacientes

A teoria de filas tem tido um papel de destaque nas pesquisas relacionadas à otimização do fluxo de pacientes. Wang *et al.* (2013, pág. 341) resumem a simplicidade e a eficiência da teoria das filas como, "*Embora os métodos analíticos contenham menos detalhes do que a simulação e sejam baseados em modelos simplificados, eles podem fornecer resultados rápidos e uma oportunidade de investigar as propriedades do sistema com mais eficiência sob suposições apropriadas*". Fomundam e Herrmann (2007) e Lakshmi e Iyer (2013) revisaram as aplicações da teoria de filas na área da saúde, desde um único departamento até o nível regional de saúde. Wiler *et al.* (2011) analisam as aplicações de modelagens relativas ao fluxo e aglomeração de pacientes no pronto-socorro, com uma seção dedicada à teoria de filas. Green (2006) detalha alguns modelos básicos de filas com aplicação em saúde, como M/M/s, M/G/1 e G/G/s.

Iniciando com algumas extensões interessantes da fila M/M/s usadas em aplicações de fluxo de pacientes em geral, não necessariamente pronto-socorro, onde uma fila com *s* servidores segue uma distribuição de chegada de *Poisson* e uma distribuição exponencial de atendimento. Duas redes de filas interativas são configuradas em Yankovic e Green (2011) sem bloqueio ou recusa de atendimento. Um sistema de filas fechadas (*CQN*) é usado em Véricourt e Jennings (2011) para avaliar a necessidade de recursos humanos, onde há uma população finita de *n* pacientes dentro do modelo M/M/s/n. Singer e Donoso (2008) usam uma aproximação M/G/s para calcular os principais indicadores de desempenho e qualidade em serviços de ambulância. Broyles e Cochran (2007) utilizaram um M/M/1/K para medir o impacto financeiro da desistência do paciente pelo serviço. A hesitação do paciente com um M/M/1/K também é vista em Cochran e Broyles (2010). Eles apontam que a precisão aliada aos dados necessários para um modelo de filas o torna um método preferido para modelagem ao invés da regressão. Um M/M/s/K aborda o fluxo de pacientes em Roche e Cochran (2007), onde usam a taxa de chegada, o tempo de serviço e o nível de utilização desejado para calcular o número de leitos *s* necessário. O uso de M/M/s no planejamento de leitos em todo o hospital é detalhado em Green e Nguyen (2001) e Green (2002), e dentro da UTI em Ridge *et al.* (1998) e Kim *et al.* (1999). Extensões de M/M/s também são vistas em Green *et al.* (2001), Green *et al.* (2006) e Green *et al.* (2007), onde as chegadas dependem do tempo e são baseadas em uma abordagem SIPP (*Stationary Independent Period by Period*). Au *et al.* (2009) utilizam uma média móvel de seis horas para representar a dependência do tempo em um modelo M/M/s para prever *overflow*. Au-Yeung *et al.* (2007) desenvolvem um modelo de filas com uma análise aproximada da função geradora, projetado para acomodar um espaço de estado maior do que os modelos tradicionais. Isso é modelado com uma rede de filas M/M/s onde os pacientes são identificados por chegada e acuidade.

As filas com servidores infinitos também são utilizadas para otimizar o fluxo de pacientes. Por exemplo, um modelo M/G/∞ é usado para analisar a alocação de leitos hospitalares e minimizar *overflow* em Kao e Tung (1981). Em última análise, a adequação do tipo de modelo de filas usado depende (i) do objetivo do estudo e (ii) das suposições subjacentes feitas. Segundo Saghafian *et al.* (2015), foram observadas quatro deficiências principais nos modelos típicos da teoria de filas utilizados para otimizar o fluxo de pacientes no pronto-socorro: (i) ignorar problemas de bloqueio/recusa no pronto-socorro; (ii) assumir chegada e processos de atendimento estacionários enquanto esses processos são de fato não estacionários; (iii) ignorar as questões de abandono no pronto-socorro e; (iv) assumir processos de

atendimento independentes do estado. Além disso, os elementos comportamentais humanos da prestação de serviços também são amplamente ignorados nos modelos de filas atuais. A seguir, uma visão geral dos artigos que tentam abordar essas deficiências.

Alguns artigos reconheceram que a fila de pacientes precisa ser modelada como uma fila finita com bloqueio para ser mais robusta (Cochran e Bharti, 2006; Osório e Bierlaire, 2009). Koizumi *et al.* (2005) incorporam o bloqueio em um modelo de filas e estendem o trabalho de um modelo tradicional de servidor único, modelando com uma rede multi-servidor M/M/s. Bretthauer *et al.* (2011) apontam falhas em modelos de bloqueio anteriores, porém, criticando o uso de uma fila com capacidade infinita. Uma heurística é utilizada para prever probabilidades de bloqueio das quais a capacidade ótima pode ser derivada.

Uma parte importante do fluxo de saída do pronto-socorro é o fenômeno *bed-block*, que se refere a situações em que os pacientes do pronto-socorro que precisam ser internados, mas não conseguem ser transferidos para as unidades de internação devido à falta de disponibilidade de leitos. Isso impede que os prontos-socorros atendam novos pacientes em tempo hábil e colabora para um maior tempo de permanência (*LOS*), bem como com uma parcela de pacientes que deixam o serviço sem serem vistos (*LWBS*). Alguns estudos de *OR/OM*, como Saghafian *et al.* (2012) consideram o efeito do fenômeno *bed-block* em várias técnicas de otimização do fluxo do paciente, incluindo "transmissão virtual". Shi *et al.* (2013) fornece um estudo detalhado de questões relacionadas à alteração dos tempos de alta nas unidades de internação, o que afeta diretamente a duração dos bloqueios de leito no pronto-socorro.

Embora a maioria dos modelos de filas assuma a condição estacionária do tempo, também há alguns trabalhos que consideram a dependência do tempo. Armony *et al.* (2011) estudam o pronto-socorro como um modelo de fila que representa um trecho de uma rede maior de filas, ou seja, todo o hospital. Três modelos de filas *Markovianos* são analisados para ajustar a ocupação do pronto-socorro junto com uma estrutura de simulação. Armony *et al.* (2011) discutem que o modelo dependente do tempo (ou seja, não estacionário), $M_t/M_t/\infty$, só é preciso quando o pronto-socorro está ocupado com menos de 15 pacientes. O modelo dependente do estado, $M_i/M_i/\infty$, em Armony *et al.* (2011) é considerado muito preciso para um modelo de filas. Esses resultados são intuitivos, pois o modelo dependente do estado incorpora fatores importantes além do tempo, como o desvio de ambulâncias. Armony *et al.* (2015) sugerem modelar a ocupação do pronto-socorro com um modelo de estado estacionário de nascimento e morte para a "caixa preta" de Dong e Whitt (2014). Um M/M/∞ também é utilizado em de Bruin *et al.* (2007) para modelar o fluxo de pacientes cardíacos da

emergência. A dependência do tempo também é capturada na modelagem de enfermarias clínicas, onde se reconhece que entender a variabilidade fora do pronto-socorro é essencial para o planejamento da capacidade (Bekker e de Bruin, 2009).

Wiler *et al.* (2013) incorporam o abandono de pacientes utilizando um modelo M/GI/r/s + GI introduzido por Whitt (2005), onde a chegada do paciente segue uma distribuição de *Poisson*, os tempos de atendimento seguem uma distribuição geral independente e identicamente distribuída, com *r* servidores (leitos), *s* capacidade da área de espera e os tempos de abandono são uma distribuição geral independente e identicamente distribuída. Batt e Terwiesch (2013) apresentam um trabalho inovador na espera do paciente, observando que fatores além do tempo de espera afetam o abandono. Entre eles estão o número observado de pacientes aguardando, o fluxo de entrada e saída de pacientes e a gravidade inferida dos pacientes em espera – todas as informações visuais adquiridas por um paciente na sala de espera.

Cochran e Roche (2009) abordam complexidades que muitas vezes são ignoradas em modelos de filas, levando em consideração a acuidade do paciente, variação da chegada e consumo de recursos em diferentes prontos-socorros. Isso também é visto em Roche e Cochran (2007) para testar um fluxo rápido no pronto-socorro. A teoria de filas também foi aplicada como uma extensão do Problema de Localização de Máxima Disponibilidade (MALP), onde é utilizada para relaxar a suposição de que a disponibilidade do servidor é aleatória (Marianov e ReVelle, 1996; Ghani, 2012). Huang (2013) e Huang *et al.* (2013) dividem os pacientes em duas redes de filas, novos pacientes e pacientes em atendimento (*WIP*), para otimizar as decisões do médico sobre quais pacientes atender. Gallivan *et al.* (2002) simplificam o processo de fluxo de pacientes por meio de um sistema determinístico de tráfego intenso, assumindo que o número de pacientes por dia, a probabilidade de "sucesso" e a permanência do paciente são sempre os mesmos.

Recentemente, um interessante estudo de simulação de alocação de leitos para reduzir a probabilidade de bloqueio nos serviços de urgência foi realizado por Wu *et al.* (2019) usando um estudo de caso na China. Eles demonstraram que as probabilidades de bloqueio podem ser significativamente reduzidas em diferentes casos de atribuição de prioridade e número total de leitos. Folake *et al.* (2020) analisaram o uso de modelos de filas na área de saúde com ênfase no pronto-socorro de um hospital municipal. Eles determinaram a contagem ideal de leitos e sua medida de desempenho para melhorar o fluxo de pacientes.

Elalouf e Wachtel (2016) desenvolveram um algoritmo que visa otimizar o agendamento de exames de pacientes, assumindo uma restrição no LOS máximo permitido no pronto-socorro. No ano seguinte (Elalouf e Wachtel, 2017), eles incorporaram uma abordagem holística ao método dinâmico de alocação de pacientes, considerando a lotação no pronto-socorro e em outros departamentos do hospital. Consideraram também a disponibilização da informação sobre a condição do paciente, além de outros fatores, tais como a gravidade e o efeito da lotação no tempo de tratamento. Ao contrário de outros domínios, o da atribuição dinâmica de recursos é relativamente novo nos prontos-socorros, uma vez que a investigação nesta área só ganhou impulso na última década.

Em 2019, Ding *et al.* (2019) analisaram as condutas dos decisores para encaminhamento dos pacientes nos quatro prontos-socorros da região metropolitana de Vancouver. Eles propuseram uma estrutura geral de escolha discreta, consistente com a literatura sobre filas, como uma ferramenta para analisar conduta de priorização em filas multi-classe. Eles observaram que as decisões, nos 4 prontos-socorros, adotavam uma abordagem de priorização (dinâmica) dependente do atraso em diferentes níveis de triagem. No mesmo ano, Zhang *et al.* (2019) propuseram um novo modelo de fila de pacientes com peso na prioridade, para otimizar o gerenciamento do pronto-socorro e, analisar o impacto desta no sistema de filas de pacientes ambulatoriais com recursos médicos limitados. Desde que foi introduzido formalmente, o conceito de agrupamento de pacientes mudou drasticamente, proporcionando um melhor atendimento ao paciente e ao gerenciamento da carga de trabalho.

Fitzgerald *et al.* (2017) usaram uma análise de Monte Carlo baseada em filas para apoiar a tomada de decisão na implementação de um processo acelerado (*fast-track*) no pronto-socorro. Eles expandiram o modelo de filas simples através de uma simulação de eventos discretos (DES) o que lhes permitiu calcular os tempos de espera. Os seus resultados indicaram que a implementação pode reduzir o tempo de espera dos pacientes sem aumentar os recursos da enfermagem. Ou seja, o *fast-track* é um dos métodos mais intuitivos para reduzir o LOS médio no pronto-socorro e, portanto, tem sido bem estudado ao longo dos anos desde o final da década de 1980, e sido muito eficaz em termos de tempo de permanência e satisfação do paciente.

Definir um conjunto de medidas de desempenho que podem capturar melhor o resultado primário é importante para avaliar quaisquer intervenções e decisões operacionais. Existem inúmeras maneiras de escolher medidas de desempenho apropriadas para um pronto-socorro

(Welch *et al.*, 2011). Na Tabela 3 estão relacionadas as medidas de desempenho mais utilizadas nos artigos sobre teoria de filas pesquisados.

Medidas de desempenho		**Artigos**	**Abordagem**
Tempo	Tempo estimado de espera	Haussmann (1970)	Relação entre as filas prioritárias e os tempos de espera dos pacientes (M/M/c)
		Taylor e Templeton (1980)	Estratégia de serviço para a qual os leitos são reservados para pacientes de alta prioridade
		Cochran e Roche (2009)	Divisão do fluxo de pacientes por acuidade ou por função (M/G/c/c)
		Madsen e Kofoed-Enevoldsen (2011)	Medidas de desempenho (M/M/1)
		Broyles e Cochran (2011)	Tempo para embarque (M/M/c)
		Silberholz *et al.* (2013)	Relação entre modelo de ensino da residência e a eficiência operacional (M/G/c)
		Lin *et al.* (2014)	Gestão do processo de saída (G/G/c)
		Saghafian *et al.* (2012, 2014)	Gestão do fluxo de pacientes/fila prioritária (G/G/c e processo de Markov)
		Almehdawe *et al.* (2013)	Simulação para validar modelos QT (processo de Markov)
		Sharif *et al.* (2014)	Simulação para validar distribuição dos tempos de espera (M/M/c)
		Yom-Tov e Mandelbaum (2014)	Simulação para validar modelos QT em sistemas grandes e pequenos para identificar inadequação
		Vass e Szabo (2015)	Medidas de desempenho (M/M/c)
		Komashie *et al.* (2015)	Nível de satisfação de pacientes e funcionários (M/G/1)
	Tempo de permanência (*LOS*)	Siddharthan e Jones (1996)	Gestão do fluxo de pacientes/fila prioritária
		de Bruin *et al.* (2007)	Estratégia para otimizar a alocação de leitos (M/M/∞, M/M/c/c)
		Mayhew e Smith (2008)	Modelo de filas para avaliar meta (NHS) de alta ≤ 4h para 98% dos pacientes
		Zeltyn *et al.* (2011)	Alocação de recursos humanos (M/M/c)
		Mandelbaum *et al.* (2012)	Sistema de filas em forma de V invertido (redução do *LOH* como estratégia para redução do tempo de embarque)
		Saghafian *et al.* (2012, 2014)	Gestão do fluxo de pacientes/fila prioritária (G/G/c e processo de Markov)
	Hora prevista de embarque (*boarding*)	Broyles e Cochran (2011)	Tempo para embarque (M/M/c)
		Lin *et al.* (2014)	Gestão do processo de saída (G/G/c)
	Fração de tempo no desvio	Allon *et al.* (2013)	Rede de filas de duas estações para modelar o fluxo de pacientes no PS ($M/M/(N_1-B)$) e na UI ($M/M/N_2/K$)
Fila	Comprimento médio da fila	Yankovic e Green (2011)	Alocação de recursos humanos (M/M/c e processo de Markov)
		Madsen e Kofoed-Enevoldsen (2011)	Medidas de desempenho (M/M/1)
		Silberholz *et al.* (2013)	Relação entre modelo de ensino da residência e a eficiência operacional (M/G/c)
		Almehdawe *et al.* (2013)	Simulação para validar modelos QT (processo de Markov)
		Zonderland *et al.* (2015)	Gestão do processo de saída
		Vass e Szabo (2015)	Medidas de desempenho (M/M/c)
	Taxa de abandono/	Green *et al.* (2006)	Redução taxa *LWBS* por médico (M/M/c)
		Cochran e Broyles (2010)	Relação entre a taxa *LWBS* e a probabilidade de recusa (M/M/1/k)

	evasão (*LWBS*)	Wiler *et al.* (2013)	Medida de *LWBS*
Probabilidade	Probabilidade de espera	de Vericourt e Jennings (2008)	Proporção enfermeiro/paciente (M/M/c/∞/n)
		Maman (2009)	Cálculo dos níveis ótimos de RH sob uma probabilidade de espera pré-especificada (M/M/c +G)
		Zeltyn *et al.* (2011)	Cálculo do RH com demanda variável no tempo (M/M/c)
		Izady e Worthington (2012)	Estimativa do %pacientes que recebem alta em até 4h para uma probabilidade de espera alvo ($M_t/G/c_t$)
		Allon *et al.* (2013)	Rede de filas de duas estações para modelar o fluxo de pacientes no PS ($M/M/(N_1-B)$) e na UI ($M/M/N_2/K$)
		Yom-Tov e Mandelbaum (2014)	Simulação para validar modelos QT em sistemas grandes e pequenos para identificar inadequação
	Probabilidade de *overflow* de área	Taylor e Templeton (1980)	Estratégia de serviço para a qual os leitos são reservados para pacientes de alta prioridade
		Au *et al.* (2009)	Gestão do processo de saída (processo de Markov)
		Cochran e Roche (2009)	Divisão do fluxo de pacientes por acuidade ou por função (M/G/c/c)
	Probabilidade de bloqueio para unidade de internação	Lin *et al.* (2014)	Gestão do processo de saída (G/G/c)
	Probabilidade de eventos adversos	Saghafian *et al.* (2014)	Gestão do fluxo de pacientes/fila prioritária (G/G/c e processo de Markov)
Recurso	Utilização de recursos	de Bruin *et al.* (2007)	Estratégia para otimizar a alocação de leitos (M/M/∞, M/M/c/c)
		Zeltyn *et al.* (2011)	Alocação de recursos humanos (M/M/c)
		Mandelbaum *et al.* (2012)	Sistema de filas em forma de V invertido (redução do *LOH* como estratégia para redução do tempo de embarque)
		Yom-Tov e Mandelbaum (2014)	Simulação para validar modelos QT em sistemas grandes e pequenos para identificar inadequação
	Requisitos de recursos adicionais	Palvannan e Teow (2012)	Como a coorte de pacientes afeta o tempo de espera de admissão no pronto-socorro (M/M/c)

Tabela 3 – Medidas de desempenho de prontos-socorros

A partir da Tabela 3, é possível observar que o tempo estimado de espera tem sido a medida mais utilizada para avaliar o desempenho de prontos-socorros, seguido pelo tempo de permanência (*LOS*) e pela probabilidade de espera. Vale ressaltar que um mesmo trabalho pode utilizar várias medidas para avaliar de forma independente ou conjunta o desempenho do serviço. Por exemplo, Saghafian *et al.* (2012) utiliza a média ponderada de *LOS* (para pacientes com alta) e tempo estimado de espera (para pacientes internados) para medir a eficácia dos vários modelos de fluxo de pacientes. Também é digno de nota que, em um sistema de filas de pronto-socorro, otimizar a pontualidade do serviço – melhor refletido pelos tempos de espera do paciente ou taxas de evasão – e a utilização de recursos (por exemplo, médicos, enfermeiros

e locais de cuidado) são objetivos conflitantes. Provedores e administradores de prontos-socorros estão constantemente tentando equilibrar o compromisso entre esses dois objetivos.

Trabalhos mais recentes com aplicação de modelos baseados na teoria de filas, como Chowdhury *et al.* (2018), Zhang A. *et al.* (2019), Hijry e Olawoyin (2022), Qandeel *et al.* (2023) e Vaghani *et al.* (2024) propõem soluções que busquem maximizar a utilização de recursos limitados sem adicionar custos significativos e adaptação de processos e recursos para horários e condições de alta demanda, objetivando reduções significativas em tempos de espera, abandonos e maior satisfação dos pacientes.

3. O SERVIÇO DE URGÊNCIA COMO UMA REDE DE FILAS

Este capítulo mostra como os serviços de urgência podem ser representados por redes de filas abertas (*OQN*). Inicialmente, apresenta-se o estado anterior ao da intervenção (*as-was*) da pesquisa-ação, principais indicadores operacionais e o fluxo de pacientes.

O pronto-socorro do hospital "H", atende diariamente mais de 300 pacientes, distribuídos nas especialidades de clínica médica (40%), ortopedia (20%), cirurgia geral (20%) e outras (20%), sendo considerado um serviço de urgência de grande porte (mais de 100 mil pacientes/ano, de acordo com *Emergency Department Benchmarking Alliance – EDBA*).

A jornada de qualquer paciente no pronto-socorro é dividida basicamente em três momentos muito bem marcados: (i) porta-médico (entrada), (ii) médico-decisão (passagem) e (iii) decisão-saída (saída), conforme ilustrado na Figura 10.

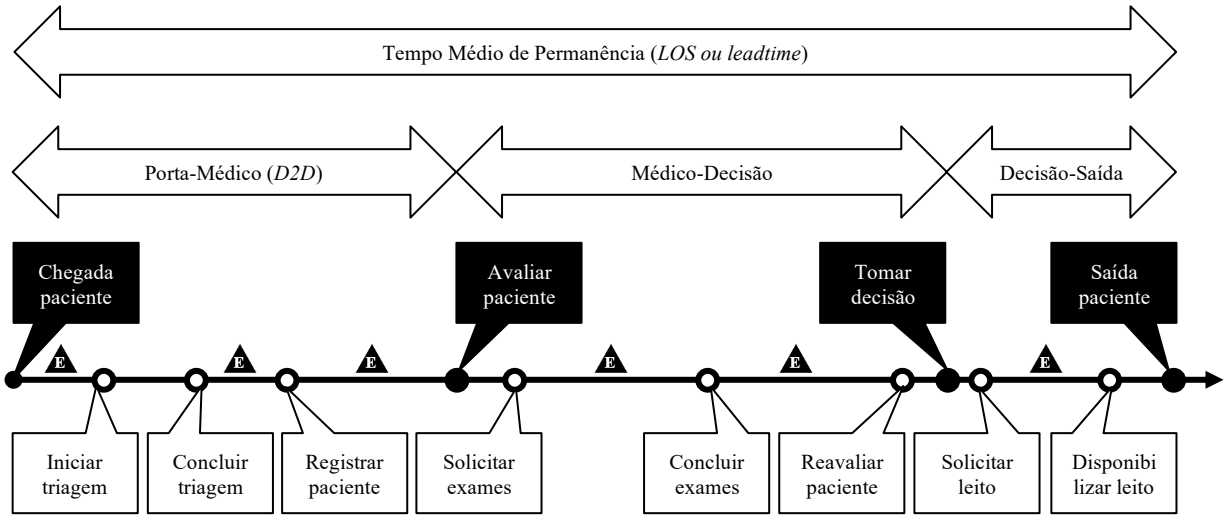


Figura 10 – Jornada do paciente no pronto-socorro (fonte: Autor)

Portanto, a jornada do paciente no pronto-socorro do "H" se inicia com a classificação de risco (iniciar triagem), passando pelo registro do paciente até ficar frente a frente com o médico (avaliar paciente). A instituição utiliza o *Manchester Triage System (MTS)* como protocolo padrão, reconhecido internacionalmente, para classificação de risco e, neste estudo, os pacientes de maior risco clínico, representado pelos emergentes (vermelho) e muito urgentes (laranja), somam 1,13%, os de risco moderado, representado pelos urgentes (amarelo) 12,78% e, os de menor risco, representado pelos pouco urgentes (verde) e os não urgentes (azul), somam 86,10%, conforme representado na Figura 11. Portanto, mais de 85% dos pacientes apresentam baixo risco clínico, entretanto, são todos colocados no mesmo fluxo, divididos apenas pela especialidade. Isso compromete consideravelmente o tempo porta-médico (*D2D*) e por consequência, o tempo médio de permanência (*LOS ou leadtime*).

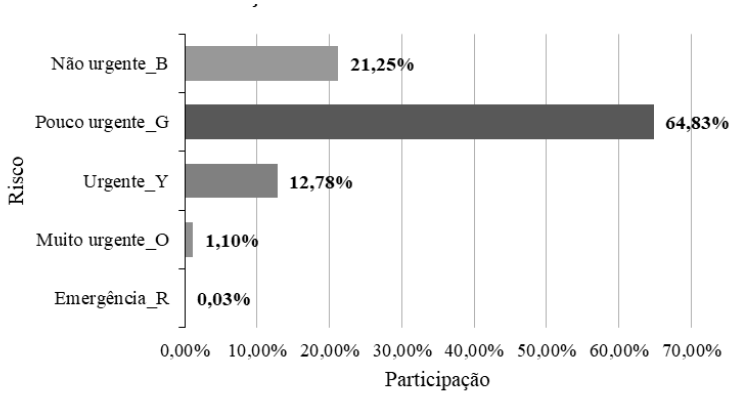


Figura 11 – Perfil de risco dos pacientes classificados no pronto-socorro conforme MTS
(fonte: Hospital "H")

O protocolo de Manchester consiste em uma triagem de classificação de risco, na qual a gravidade dos casos é determinada por cores e, portanto, mediante a colocação de uma pulseira na cor correspondente, os pacientes são identificados. A classificação do risco, pelo protocolo de Manchester, também determina o tempo de espera (priorização) para atendimento do paciente, conforme Tabela 4:

Risco	Emergência	Muito Urgente	Urgente	Pouco urgente	Não urgente
Cor	Vermelho	Laranja	Amarelo	Verde	Azul
Tempo de espera	0 min.	até 10 min.	até 50 min.	até 120 min.	até 240 min

Tabela 4 – Protocolo de Manchester

O pronto-socorro é um ambiente bastante dinâmico e, normalmente, apresenta um fluxo grande de pacientes, portanto, requer indicadores de processo que proporcionem aos seus gestores, para além da avaliação clínica, a documentação e o estabelecimento de prioridades nos serviços ofertados. A Tabela 5 apresenta os principais indicadores utilizados em serviços de urgência:

Indicador	**Descrição**
Tempo porta-médico (*D2D*)	É o intervalo de tempo, em minutos, entre a chegada do paciente no pronto-socorro e a sua presença frente a frente com o médico. É um dos principais indicadores do pronto-socorro, uma vez que, a taxa de evasão e o índice de satisfação do paciente estão diretamente relacionados com ele. Dentro deste intervalo de tempo estão inclusos os tempos de registro e de classificação do risco do paciente, bem como as esperas entre estes processos.
Tempo médico-decisão	Corresponde ao intervalo de tempo, em minutos, entre o início da consulta médica e a decisão médica pela alta ou internação do paciente. Dentro deste intervalo estão os tempos de consulta, realização de exames (laboratório e/ou imagem) e reavaliação médica, bem como as esperas entre estes processos.
TAT (*TurnAroundTime*) Lab./Rad.	Intervalo de tempo, em minutos, entre a solicitação e o resultado (disponível para o médico) do exame de laboratório e/ou imagem (para que o médico possa tomar a sua decisão).

Tempo decisão-saída	Corresponde ao intervalo de tempo, em minutos, entre a decisão médica e a saída física do paciente do pronto-socorro. **Tempo de *boarding* –** Caso a decisão médica tenha sido pela internação do paciente, este indicador se assume como tempo de embarque (*boarding*) e, corresponde ao tempo em que o paciente permanece "internado" no pronto-socorro aguardando por um leito em uma das unidades de internação (UTI/Enfermarias) do hospital. Este tempo pode variar desde poucas horas até muitos dias e, em casos extremos, o paciente recebe alta a partir do próprio pronto-socorro, sem sequer acessar um leito de internação.
Tempo médio de permanência (*LOS ou leadtime*)	Também conhecido como *LOS* (*Lenght Of Stay*) ou *leadtime*, corresponde ao intervalo de tempo entre a chegada do paciente até a sua saída física do pronto-socorro (por alta, internação ou transferência), portanto, corresponde à somatória dos tempos porta-médico, médico-decisão e decisão-saída. É um dos mais importantes indicadores de desempenho de um pronto-socorro.
Taxa de evasão (*LWBS*)	Consiste na razão entre a quantidade de pacientes (pré-consulta), de um determinado período, que evadem do serviço e, o total de atendimentos relativo ao mesmo período. Este indicador é sensível ao tempo porta-médico (*D2D*).
Taxa de conversão	Número de internações em relação ao total de atendimentos no pronto-socorro.
Índice de satisfação do cliente (*NPS*)	É comum utilizar o *Net Promoter Score* para avaliar serviços hospitalares, dado a sua simplicidade, confiabilidade e flexibilidade. Consiste em uma métrica de lealdade do cliente, criada por Frederick Reichheld em 2003 através da publicação de um artigo chamado *The One Number You Need to Grow*, na revista da Universidade Harvard, com o objetivo de medir o grau de lealdade dos clientes das empresas, de qualquer segmento, trazendo reflexos da sua experiência e satisfação. O *NPS* é calculado com base nas respostas a uma única pergunta: "Qual é a probabilidade de você recomendar o nosso serviço a um amigo ou colega?". A pontuação para esta resposta varia com base em uma escala de 0 a 10. (i) Aqueles que respondem com uma pontuação de 9 ou 10 são chamados de Promotores, e são considerados propensos a apresentar comportamentos de criação de valor, tais como permanecer clientes por mais tempo e fazer referências positivas para outros potenciais clientes. (ii) Aqueles que respondem com uma pontuação de 0 a 6 são rotulados como Detratores, e acredita-se ser menos propensos a apresentar comportamentos de criação de valor. (iii) Respostas de 7 ou 8 são rotulados como Passivos ou Neutros, e seu comportamento cai no meio de promotores e detratores. O *NPS* é calculado subtraindo a porcentagem entre Detratores e Promotores. Clientes Passivos ou Neutros contam para o número total de entrevistados, mas não afetam diretamente o resultado líquido global.

Tabela 5 – Indicadores de processo mais utilizados em serviços de urgência (fonte: Autor)

Através da análise descritiva e diagnóstica realizada *in loco* durante 5 dias pelo autor, foram levantados uma série de dados cujos indicadores encontram-se listados na Tabela 6.

Indicadores de processo	**PS "H"**	**EDBA**
Tempo médio de permanência (*LOS*), min.	265	140
Tempo porta-médico (*D2D*), min.	65	16
Tempo médico-decisão, min.	140	-
Tempo (*TAT*) de resposta Lab./Rad., min.	120	-
Tempo decisão-saída (alta), min.	15	-
Tempo de *boarding*, min.	60	71
Taxa de evasão (*LWBS*), %	2,2	1,5
Taxa de conversão, %	6,25	11,0

Tabela 6 – Indicadores PS "H" e referência EDBA
(fonte: Dark C. *et al.*, 2020)

Comparando os indicadores do pronto-socorro "H" com a referência EDBA (Dark C. *et al.*, 2020), o tempo porta-médico (*D2D*) e a taxa de evasão (*LWBS*) foram superiores à referência. Sendo estes, os dois indicadores que mais influenciam na avaliação da experiência do usuário (*NPS*), a quantidade de detratores (clientes que tiveram uma experiência extremamente negativa com a instituição) e de neutros (clientes que tiveram uma experiência mediana com a instituição e, são vulneráveis às ofertas da concorrência) levaram a instituição a uma baixa avaliação, conforme Figura 12, o que a coloca em uma "zona de aperfeiçoamento", ou seja, é preciso realizar melhorias no(s) serviço(s) ofertado(s).

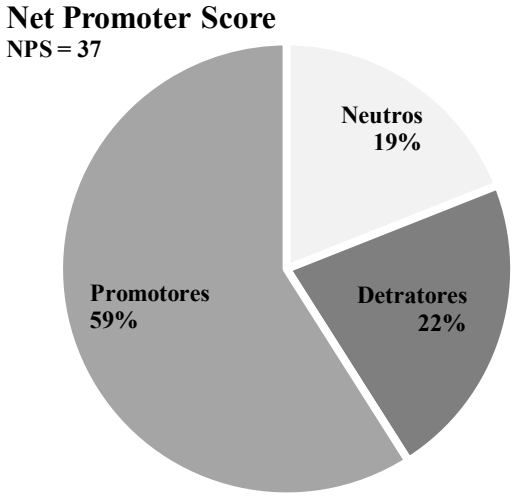


Figura 12 – *Net Promoter Score – NPS*
(fonte: Hospital "H")

O segundo momento (médico-decisão) responde relativamente bem e não compromete a jornada do paciente. Neste intervalo de tempo estão inclusos a consulta médica, a realização de exames de laboratório e/ou imagem e a reavaliação médica, conforme Tabela 6. Vale destacar que a decisão pode se dar "pela alta", que em geral demanda menos tempo ou "pela internação", neste caso, representada por 6,25% das decisões (taxa de conversão). Cerca de 30% dos pacientes realizam exames de laboratório e outros 25% exames de imagem (raio-x), cujos tempos de resposta (*TAT*), em ambos os casos, demandaram cerca de 120 minutos. Vale destacar que, elevada demanda por exames e tempos de resposta igualmente altos, podem comprometer consideravelmente o tempo médico-decisão e, por consequência, o tempo médio de permanência (*LOS*) do paciente.

O terceiro e último momento (decisão-saída), para os pacientes cuja decisão médica foi "pela alta", refere-se basicamente ao tempo necessário para percorrer os corredores do pronto-socorro e sair, tendo sido apontado 15 minutos, em média. Já aqueles pacientes cuja decisão foi "pela internação", este tempo foi de 60 minutos (tempo de *boarding*, ver Tabela 6). Tempo ótimo quando comparado com a referência EDBA na própria Tabela 6, mas que neste caso poderia ser ainda melhor em função da subutilização (ρ=47%) dos leitos nas unidades de internação.

Desta forma, o tempo médio ponderado de permanência do paciente no pronto-socorro do hospital "H" foi elevado em comparação com os hospitais referência EDBA, de igual porte, na Tabela 6. Entretanto, como o percentual de pacientes de baixo risco clínico é superior a 85%, este tempo pode ser melhorado, uma vez que este é o perfil de pacientes, que menos consome recursos do pronto-socorro.

Observando o fluxo de pacientes no pronto-socorro "H", ilustrado na Figura 13, os pacientes ao chegarem ao PS passam pela classificação de risco (*TRG*) e na sequência pelo registro (*REG*). O fluxo então se divide em três:

1) os mais graves, e que necessitam serem estabilizados $r_{TRG,VSM}^{(R)}$ são encaminhados para a sala vermelha (*SVM*);
2) os que demandam uma intervenção cirúrgica de urgência (*CRG*) $r_{TRG,CRG}^{(R,O)}$ e, por último;
3) o maior fluxo do pronto-socorro, consultório médico (*DOC*), representado por cerca de 98% do total de pacientes $r_{TRG,DOC}^{(O,Y,G,B)}$ e composto, na sua grande maioria, por pacientes de baixo risco clínico (verdes + azuis), sendo segmentados apenas pela especialidade médica (*DOC*) e não pelo risco.

Os pacientes deste último fluxo (3) podem realizar:

i. apenas uma consulta médica (*DOC*) e saírem em alta (*ALT*) do pronto-socorro ou;
ii. além da consulta médica (*DOC*), podem realizar exames de laboratório (*LAB*) e/ou exames de imagem (*RAD*) e/ou tratamento (*TRT*) (medicação, observação, entre outros procedimentos) e retornam ao médico (*DOC*) para a reavaliação e tomada de decisão.

A grande maioria destes, deixam o PS em alta (*ALT*) e, apenas uma pequena parcela $r_{TRT,INT}^{(O,Y)}$, somada àqueles provenientes do Centro Cirúrgico (CRG) ($r_{CRG,INT}^{(R,O)}$) e da Sala Vermelha (*SVM*) ($r_{SVM,INT}^{(R)}$), seguirão para as unidades de internação (*INT*).

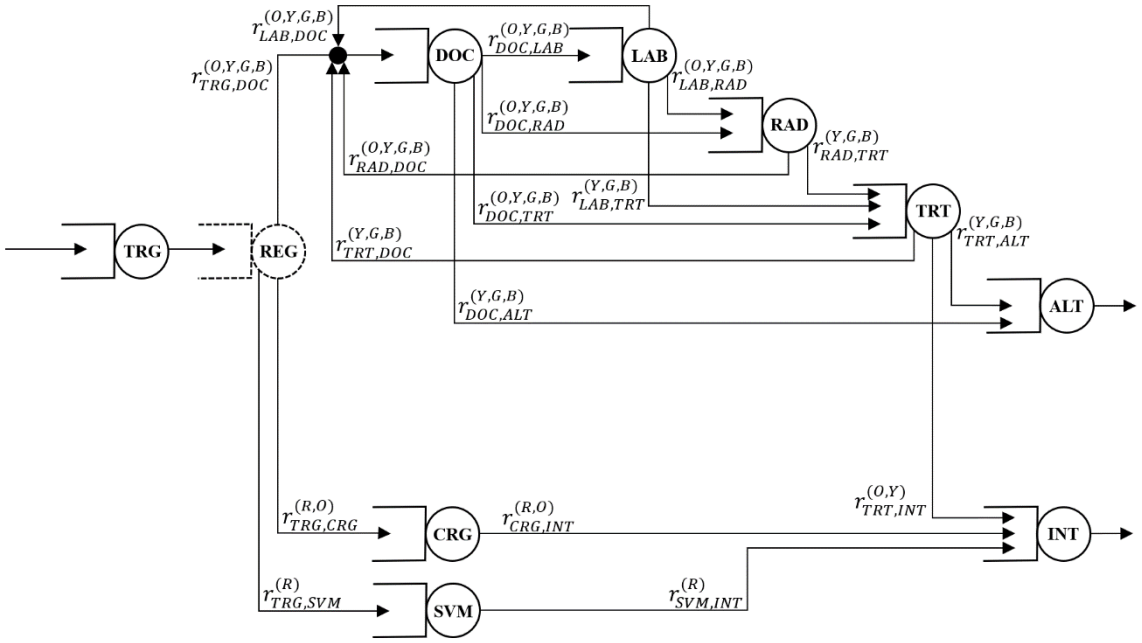


Figura 13 – Rede de filas com **priorização** (*as-was*) dos pacientes nos 9 nós do PS "H"
onde $r_{i,j}^{(t)}$ representa a probabilidade de fluxo do nó *i* para *j* conforme o risco *t* do paciente (fonte: Autor)

4. ANÁLISE DO OS UTILIZANDO REDE DE FILAS MULTI-CLASSES

Este capítulo mostra como o serviço de urgência do hospital "H" pode ser melhorado. Inicialmente apresenta-se a nova configuração da rede de filas (seção 4.1), a metodologia proposta e os resultados computacionais (seção 4.2), a geração e análise das curvas de *trade-off* (seção 4.3) e, o capítulo termina com uma análise sobre o giro de leitos nas unidades de internação (seção 4.4) afim de mitigar o efeito do *boarding* no pronto-socorro.

4.1. Configuração proposta para o pronto-socorro

O que se propõe nesta nova configuração é, a partir do protocolo de Manchester, a "segmentação" dos pacientes, ou seja, estabelecer fluxos distintos para pacientes com necessidades, inclusive de recursos, distintas, conforme ilustrado pela Figura 14.

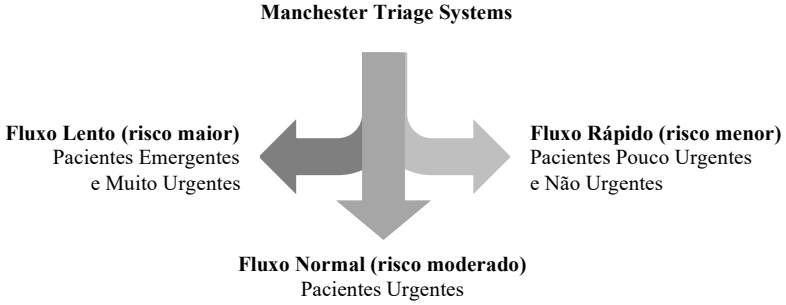


Figura 14 – Segmentação dos pacientes (fonte: Autor)

Para esta nova configuração da rede de filas do pronto-socorro do hospital "H", a proposta foi implementada com a inclusão de três novas áreas:

i. os consultórios de *fast-track* (*FTK*), para um fluxo rápido dos pacientes de baixo risco clínico $r_{TRG,FTK}^{(G,B)}$, conforme Figura 15 e, cujo objetivo é desviar boa parte daqueles pacientes $r_{TRG,DOC}^{(O,Y,G,B)}$ que inicialmente, conforme Figura 13, demandariam apenas uma consulta médica (*DOC*);

ii. o consultório de teleconsulta (*TELE*), um fluxo igualmente rápido também para pacientes de baixo risco clínico $r_{TRG,TELE}^{(G,B)}$, conforme Figura 15, podendo ser acessado de forma remota ou eventualmente presencial no próprio PS e;

iii. a unidade de decisão clínica (*UDC*), uma área dedicada a pacientes cuja decisão médica demanda mais tempo $r_{DOC,UDC}^{(O,Y)}$, pela necessidade de mais exames e/ou uma segunda opinião médica, porém, limitada em até 4 horas.

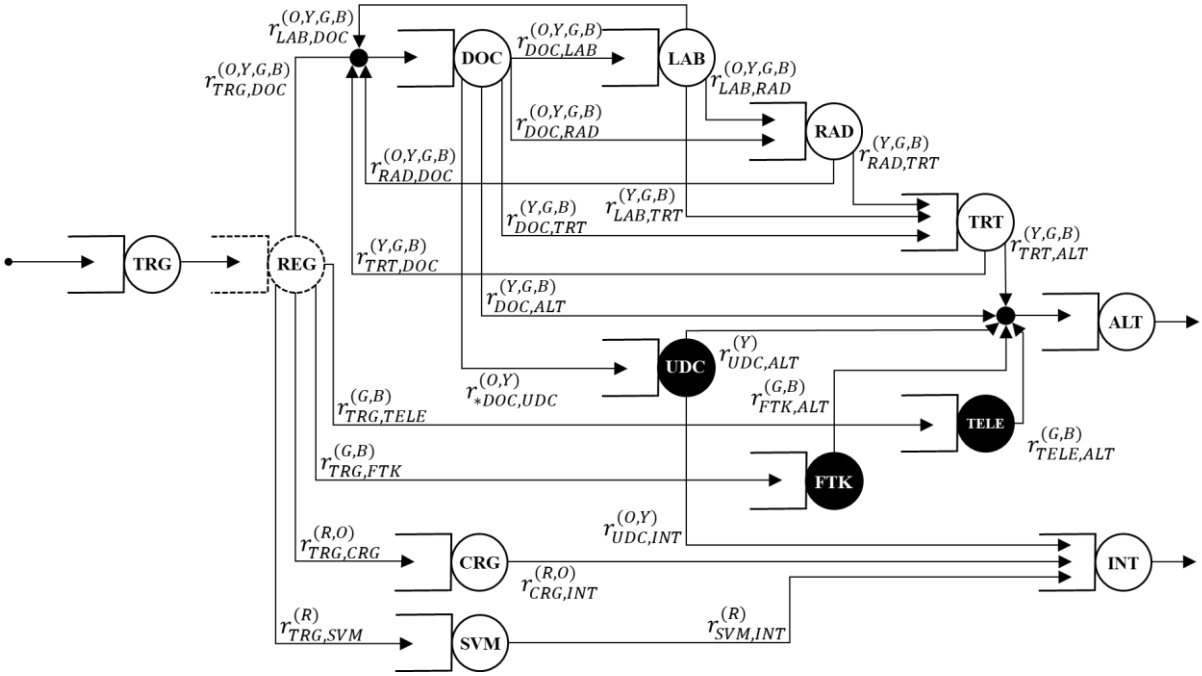


Figura 15 – Rede de filas com **segmentação** (*as-is*) dos pacientes nos 12 nós do PS "H"
onde $r_{i,j}^{(t)}$ representa a probabilidade de fluxo do nó *i* para *j* conforme o risco *t* do paciente (fonte: Autor)

Já o fluxo dos pacientes de maior risco clínico, que se destinam às áreas de Centro Cirúrgico (*CRG*) e Sala Vermelha (*SVM*), permanecem inalterados.

Com a inclusão dessas áreas e as alterações advindas dessa inclusão, houve maior fluidez no atendimento aos pacientes de baixo risco clínico, retirando-os do pronto-socorro em menor tempo. Isso libera recursos físicos e humanos, permitindo novos atendimentos de maneira mais eficiente.

4.2. Metodologia proposta

A metodologia proposta foi inspirada no trabalho de Cochran e Roche (2009). Ao adotar uma perspectiva multi-classe, ambos trabalhos destacam a capacidade de lidar com diferentes categorias de pacientes, reconhecendo que as necessidades e tempos de atendimento podem variar consideravelmente. Além disso, equações de filas foram utilizadas para dimensionar as áreas (nós) do pronto-socorro, usando o tempo de espera (W_q), o fator de utilização (ρ) e a probabilidade de superlotação (p_c) como metas de qualidade do serviço.

Enquanto o pronto-socorro do hospital de Phoenix no Arizona apresenta uma rede de 9 nós e atende cerca de 55 mil pacientes/ano, classificados através do também americano protocolo ESI, o pronto-socorro do hospital "H", apresenta uma rede de 12 nós e atende pouco mais de 116 mil pacientes/ano, classificados através do Protocolo de Manchester (MTS). Outra diferença diz respeito à aplicação, enquanto o trabalho de Cochran e Roche (2009) capacita os

nós do pronto-socorro e calcula o tempo porta-médico com base na qualidade requerida do serviço, neste, dado o número de recursos (físicos e humanos), busca-se encontrar o melhor *trade-off* de forma a atender à qualidade requerida do serviço.

Ambos trabalhos destacam a importância de abordagens quantitativas na gestão de serviços de saúde, contribuindo para a base teórica e prática da tomada de decisões em prontos-socorros. A capacidade de modelar e analisar o fluxo de pacientes em um ambiente de emergência hospitalar é crucial para a implementação de melhorias significativas.

A Figura 16 apresenta a metodologia proposta nesta dissertação para analisar a capacidade de cada uma das áreas (nós) do pronto-socorro. Cada uma das etapas é descrita nas próximas subseções.

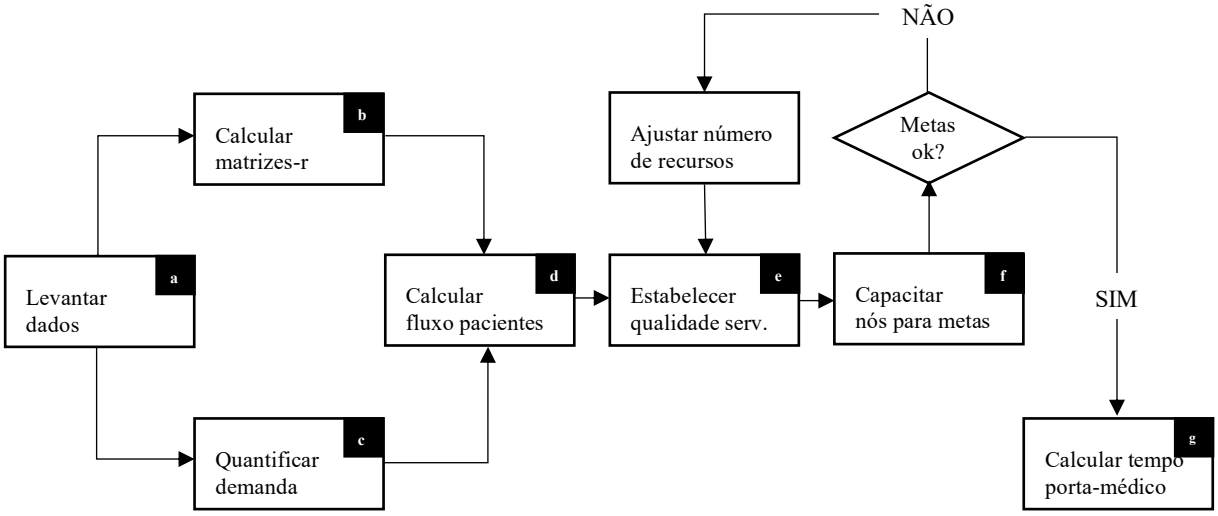


Figura 16 – Metodologia proposta para análise de capacidade das áreas (nós) do PS (fonte: Autor)

a) Coletar os dados

Para modelar com precisão o fluxo de pacientes pela rede, junto ao hospital em análise, foram levantados os dados de entrada dos pacientes, como por exemplo: a taxa média de chegada pelo pronto-socorro (Figura 17), os tempos médios de atendimento (ciclo) para cada uma das áreas (nós) e os tempos de espera entre as mesmas. O modelo proposto foi intencionalmente projetado para utilizar a forma e o conteúdo dos dados normalmente disponíveis em um pronto-socorro.

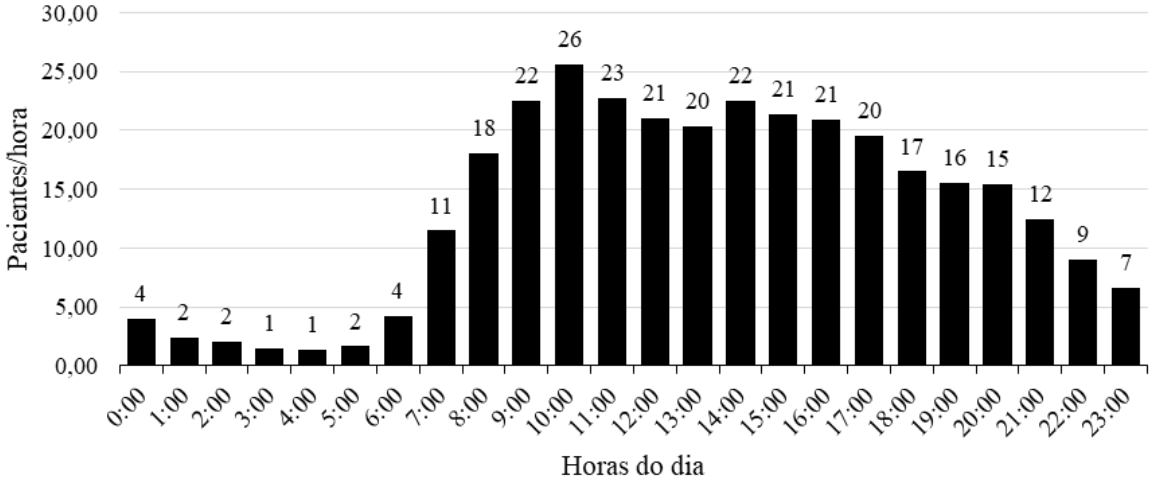


Figura 17 – Taxa média de chegada por hora no pronto-socorro (fonte: hospital "H")

Ao chegar, os pacientes são classificados, em um dos 5 níveis de risco, utilizando o protocolo de Manchester (*MTS*). Estes níveis podem ser mensurados a partir dos dados históricos do hospital, conforme Figura 11. Para calibrar o nível geral de demanda na rede de fluxo de pacientes, foi utilizado um volume anual histórico conforme Figura 17 e, a partir deste, foram considerados os efeitos de sazonalidade (mensal) e de pico (diário).

A estimativa de volume também incluiu os pacientes que deixam o serviço sem serem vistos pelo médico (*LWBS*) uma vez que esta metodologia visa melhorar a segurança do paciente e mitigar a taxa de evasão dos mesmos. Observando a Figura 18 percebe-se que nos meses em que o volume ficou próximo de 12.000 pacientes a taxa de evasão oscilou entre 3% e 4%. A taxa de evasão (*LWBS*), bem como o índice de satisfação do paciente (*NPS*) são muito sensíveis ao tempo porta-médico (*D2D*), ou seja, elevadas taxas de evasão e baixo índice de satisfação estão diretamente relacionadas com um tempo porta-médico alto, que é o caso do pronto-socorro em análise.

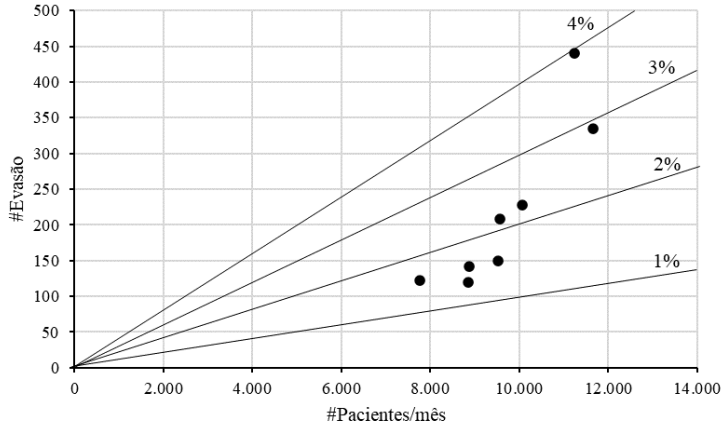


Figura 18 – Taxa de evasão (*LWBS – Left Without Be Seen*) (fonte: hospital "H")

Para contabilizar as taxas de serviço nas áreas de assistência, a mediana dos tempos de ciclo, para cada um dos nós da rede (i) foi calculada a partir dos dados extraídos do *ERP* do hospital, no período analisado. Como os hospitais normalmente têm uma compreensão razoável do tempo em que os pacientes permanecem em cada uma destas áreas, os tempos extraídos do sistema também foram validados com especialistas do hospital. O inverso do tempo de ciclo, é a taxa de serviço (μ), dada em pacientes por hora.

Os pacientes saem do serviço de urgência com base em frações de pacientes admitidos (*INT*) no hospital ou frações de pacientes em alta (*ALT*). Estes dados também estão disponíveis no sistema integrado de gestão (*ERP*) do hospital.

b) Calcular as matrizes-r

As matrizes de roteamento ou matrizes-r capturam o padrão de fluxo de cada perfil de risco dos pacientes cujas probabilidades de roteamento se encontram apenas nos pares de nós que estão diretamente conectados e, portanto, são matrizes esparsas. Por definição, quando um paciente deixa o pronto-socorro, ele está fora do conjunto de nós modelados e nenhuma probabilidade de roteamento correspondente aparece na matriz-r. Foram derivadas cinco matrizes-r, uma para cada nível de classificação de risco de pacientes do protocolo de Manchester.

As matrizes-r para pacientes cujos perfis de risco são emergência (*R*) e muito urgente (*O*), onde $i, j \in$ {Triagem (*TRG*), Consulta Médica (*DOC*), Teleconsulta (*TELE*), *Fast-Track* (*FTK*), Centro Cirúrgico (*CRG*), Sala Vermelha (*SVM*), Laboratório (*LAB*), Radiologia (*RAD*), Tratamento (*TRT*), Unidade de Decisão Clínica (*UDC*), Alta Médica (*ALT*) e Unidades de Internação (*INT*)}, encontram-se abaixo representadas:

$$\boldsymbol{r}_{\boldsymbol{ij}}^{(\boldsymbol{R})} = \begin{bmatrix} \ldots & 0 & 0 & 0 & r_{TRG,CRG} & r_{TRG,SVM} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & r_{INT,CRG} \\ 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & r_{INT,SVM} \\ 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \end{bmatrix}$$

$$\boldsymbol{r}_{\boldsymbol{ij}}^{(\boldsymbol{R})} = \begin{bmatrix} \ldots & 0 & 0 & 0 & 0{,}0385 & 0{,}9616 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0{,}0385 \\ 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0{,}9616 \\ 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \end{bmatrix}$$

$$\boldsymbol{r}_{\boldsymbol{ij}}^{(\boldsymbol{O})} = \begin{bmatrix} \ldots & r_{TRG,DOC} & 0 & 0 & r_{TRG,CRG} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \ldots & 0 & 0 & 0 & 0 & r_{DOC,LAB} & r_{DOC,RAD} & 0 & r_{DOC,UDC} & 0 & 0 \\ 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & r_{CRG,INT} \\ 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & r_{LAB,DOC} & 0 & 0 & 0 & 0 & \ldots & r_{LAB,RAD} & 0 & 0 & 0 & 0 \\ 0 & r_{RAD,DOC} & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & r_{UDC,INT} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \end{bmatrix}$$

$$\boldsymbol{r}_{\boldsymbol{ij}}^{(\boldsymbol{O})} = \begin{bmatrix} \ldots & 0{,}9696 & 0 & 0 & 0{,}0304 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \ldots & 0 & 0 & 0 & 0 & 0{,}2930 & 0{,}2000 & 0 & 0{,}9696 & 0 & 0 \\ 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0{,}0304 \\ 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0{,}2930 & 0 & 0 & 0 & 0 & \ldots & 0{,}0310 & 0 & 0 & 0 & 0 \\ 0 & 0{,}2310 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0{,}9696 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \end{bmatrix}$$

A matriz-r para os pacientes cujo perfil é urgente (*Y*), apresenta mais conexões e, por consequência, menos elementos nulos.

$$\boldsymbol{r}_{\boldsymbol{ij}}^{(\boldsymbol{Y})} = \begin{bmatrix} \ldots & r_{TRG,DOC} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \ldots & 0 & 0 & 0 & 0 & r_{DOC,LAB} & r_{DOC,RAD} & r_{DOC,TRT} & r_{DOC,UDC} & r_{DOC,ALT} & 0 \\ 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & r_{LAB,DOC} & 0 & 0 & 0 & 0 & \ldots & r_{LAB,RAD} & r_{LAB,TRT} & 0 & 0 & 0 \\ 0 & r_{RAD,DOC} & 0 & 0 & 0 & 0 & 0 & \ldots & r_{RAD,TRT} & 0 & 0 & 0 \\ 0 & r_{TRT,DOC} & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & r_{TRT,ALT} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & r_{UDC,ALT} & r_{UDC,INT} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \end{bmatrix}$$

$$\boldsymbol{r}_{\boldsymbol{ij}}^{(\boldsymbol{Y})} = \begin{bmatrix} \ldots & 1{,}0000 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \ldots & 0 & 0 & 0 & 0 & 0{,}2930 & 0{,}2000 & 0{,}1070 & 0{,}2000 & 0{,}2000 & 0 \\ 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0{,}2680 & 0 & 0 & 0 & 0 & \ldots & 0{,}0310 & 0{,}0250 & 0 & 0 & 0 \\ 0 & 0{,}2060 & 0 & 0 & 0 & 0 & 0 & \ldots & 0{,}0250 & 0 & 0 & 0 \\ 0 & 0{,}0100 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0{,}1000 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0{,}2000 & 0{,}4000 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \end{bmatrix}$$

As probabilidades de segmentação dos perfis de risco, pouco urgente (*G*) e não urgente (*B*) estão apresentadas nas matrizes abaixo:

$$\boldsymbol{r}_{\boldsymbol{ij}}^{(\boldsymbol{G})} = \begin{bmatrix} \ldots & r_{TRG,DOC} & 0r_{TRG,TELE} & r_{TRG,FTK} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \ldots & 0 & 0 & 0 & 0 & r_{DOC,LAB} & r_{DOC,RAD} & r_{DOC,TRT} & 0 & r_{DOC,ALT} & 0 \\ 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & r_{TELE,ALT} & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & r_{FTK,ALT} & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & r_{LAB,DOC} & 0 & 0 & 0 & 0 & \ldots & r_{LAB,RAD} & r_{LAB,TRT} & 0 & 0 & 0 \\ 0 & r_{RAD,LAB} & 0 & 0 & 0 & 0 & 0 & \ldots & r_{RAD,TRT} & 0 & 0 & 0 \\ 0 & r_{TRT,LAB} & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & r_{TRT,ALT} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \end{bmatrix}$$

$$\boldsymbol{r}_{\boldsymbol{ij}}^{(\boldsymbol{G})} = \begin{bmatrix} \ldots & 0{,}5000 & 0{,}1000 & 0{,}4000 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \ldots & 0 & 0 & 0 & 0 & 0{,}1465 & 0{,}1000 & 0{,}1000 & 0 & 0{,}3500 & 0 \\ 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0{,}1000 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0{,}4000 & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0{,}1215 & 0 & 0 & 0 & 0 & \ldots & 0{,}0155 & 0{,}0250 & 0 & 0 & 0 \\ 0 & 0{,}0905 & 0 & 0 & 0 & 0 & 0 & \ldots & 0{,}0250 & 0 & 0 & 0 \\ 0 & 0{,}0050 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0{,}1500 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \end{bmatrix}$$

$$\boldsymbol{r}_{\boldsymbol{ij}}^{(\boldsymbol{B})} = \begin{bmatrix} \ldots & r_{TRG,DOC} & r_{TRG,TELE} & r_{TRG,FTK} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \ldots & 0 & 0 & 0 & 0 & r_{DOC,LAB} & r_{DOC,RAD} & r_{DOC,TRT} & 0 & r_{DOC,ALT} & 0 \\ 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & r_{TELE,ALT} & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & r_{FTK,ALT} & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & r_{LAB,DOC} & 0 & 0 & 0 & 0 & \ldots & r_{LAB,RAD} & r_{LAB,TRT} & 0 & 0 & 0 \\ 0 & r_{RAD,LAB} & 0 & 0 & 0 & 0 & 0 & \ldots & r_{RAD,TRT} & 0 & 0 & 0 \\ 0 & r_{TRT,LAB} & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & r_{TRT,ALT} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \end{bmatrix}$$

$$\boldsymbol{r}_{\boldsymbol{ij}}^{(\boldsymbol{B})} = \begin{bmatrix} \ldots & 0{,}5000 & 0{,}1000 & 0{,}4000 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \ldots & 0 & 0 & 0 & 0 & 0{,}1465 & 0{,}1000 & 0{,}1000 & 0 & 0{,}3500 & 0 \\ 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0{,}1000 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0{,}4000 & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0{,}1205 & 0 & 0 & 0 & 0 & \ldots & 0{,}0155 & 0{,}0250 & 0 & 0 & 0 \\ 0 & 0{,}0905 & 0 & 0 & 0 & 0 & 0 & \ldots & 0{,}0250 & 0 & 0 & 0 \\ 0 & 0{,}0050 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0{,}1500 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \end{bmatrix}$$

Pacientes de maior risco clínico (nível emergência e muito urgente) se conectam com a sala vermelha (*SVM*), o centro cirúrgico (*CRG*) e apenas uma pequena parcela dos muito urgentes segue o fluxo tradicional (*DOC*). Os pacientes de risco clínico moderado (urgente) transitam e se conectam com praticamente todas as áreas do pronto-socorro. Já os pacientes de menor risco clínico (pouco urgentes e não urgentes) são, na sua maioria, direcionados para o *fast-track*, ou a teleconsulta, áreas, idealmente posicionadas próxima à porta de entrada do pronto-socorro para que os pacientes, após sua consulta, possam sair por esta mesma porta, evitando circulação desnecessária pelas dependências do pronto-socorro. A área de teleconsulta pode, inclusive, ser acessada de forma remota, evitando o deslocamento do paciente até o pronto-socorro.

c) Quantificar a demanda

As chegadas ao pronto-socorro não são homogêneas no tempo. Os padrões de chegada de pacientes por hora e as sazonalidades devem ser considerados ao prover capacidade ao pronto-socorro. A sazonalidade no dimensionamento de prontos-socorros é vista como subjetiva e dependente das características de cada instituição em análise. A variação sazonal pode ser modelada usando índices multiplicadores, como aqueles apresentados em Ozcan (2005). O erro comum é usar a média ao longo do ano. Outros pesquisadores na área da saúde observaram um padrão consistente de chegada de pacientes ao pronto-socorro durante o dia. Em qualquer caso, a Equação (1) é uma fórmula geral para converter um volume anual da emergência em chegadas de pacientes por hora usando picos horários (Figura 19) e/ou sazonalidades (Figura 20):

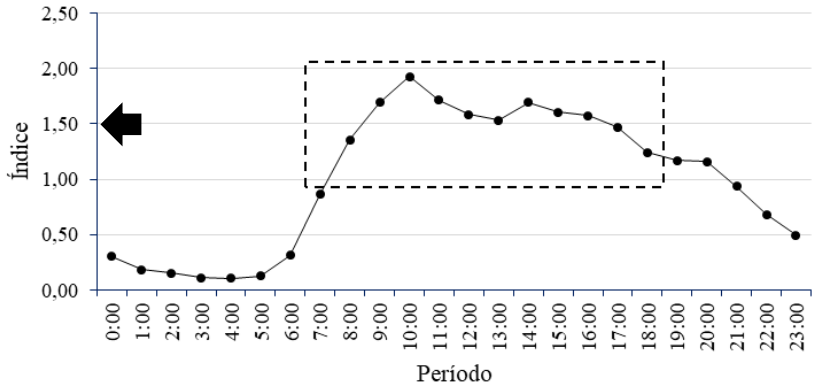


Figura 19 – Índice multiplicador (efeitos dos picos) para chegada diária (fonte: Autor)
Média #pacientes/hora no período de maior demanda (7 às 19) dividido pela média global de pacientes/dia (319 pts./dia) x 24

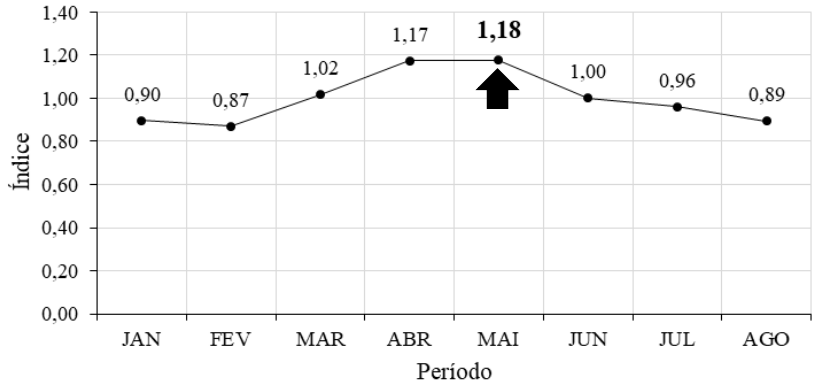


Figura 20 – Índice multiplicador (efeito sazonalidade) para chegada mensal (fonte: Autor)
Média #pacientes/dia por mês do período analisado dividido pela média global de pacientes/dia (319 pts./dia)

Somado o número de pacientes/hora da Figura 17 e multiplicado por 365 dias/ano tem-se o numerador da Equação (1):

$$\lambda_i = \frac{\frac{\#\text{pacientes}}{\text{ano}}}{365\frac{\text{dias}}{\text{ano}} \times 24\frac{\text{horas}}{\text{dia}}} \times \text{Índice}_{\text{sazonalidade}} \times \text{Índice}_{\text{picos}} \quad (1)$$

$$\lambda_i = \frac{116.274\frac{\text{pts}}{\text{ano}}}{365\frac{\text{dias}}{\text{ano}} \times 24\frac{\text{horas}}{\text{dia}}} \times 1{,}18 \times 1{,}50 = \mathbf{23{,}49}\,\frac{\mathbf{pts}}{\mathbf{hora}}$$

d) Calcular o fluxo de pacientes nos nós

A segmentação do fluxo de pacientes no pronto-socorro é tratada como uma rede aberta de filas (*OQN*). Usando a notação de Gross e Harris (1998), a taxa de chegada total a um nó em uma rede de filas aberta é calculada utilizando-se as matrizes-r e as taxas de chegada:

$$\lambda_i = \sum_{t=1}^{T} \gamma_i^{(t)} + \sum_{t=1}^{T} \sum_{j=1}^{n} \lambda_j^{(t)} \times r_{ji}^{(t)} \quad (2)$$

Onde:

λ_i é a taxa total de chegada ao nó *i*;

$\gamma_i^{(t)}$ é a taxa de chegada externa até o nó *i* do tipo de paciente *t* e;

$\sum_{t=1}^{T} \sum_{j=1}^{n} \lambda_j^{(t)} \times r_{ji}^{(t)}$ representa a taxa de chegada de pacientes transferidos de todos os outros nós para o nó *i* para todos os tipos de pacientes *T*.

No modelo de segmentação de pacientes do pronto-socorro, a única taxa de chegada externa ocorre no nó triagem (*TRG*), calculada na etapa "c" (quantificar a demanda). As taxas de chegada aos outros nós são limitadas às transferências internas entre os nós. Portanto, a Equação (1) pode ser simplificada pela Equação (3):

$$\lambda_i = \sum_{t=1}^{5} \sum_{j=1}^{12} \lambda_j^{(t)} \times r_{ji}^{(t)} \quad (3)$$

Todos os termos $r_{ji}^{(t)}$ foram definidos nas matrizes-r calculadas na etapa "b" da metodologia para cada um dos cinco perfis de pacientes. Este conjunto de equações é, em geral e particularmente, fácil de resolver nesta aplicação, pois há apenas fluxo direto, portanto, as equações podem ser resolvidas sequencialmente em vez de simultaneamente, conforme demonstrado na Tabela 7.

$\boldsymbol{\lambda_i}$	**Pts./h**
$\lambda_{TRG} = 23{,}4937 \times (1{,}0000) =$	23,4937
$\lambda_{DOC} = 23{,}4937 \times (0{,}0110 \times 1{,}4936 + 0{,}1278 \times 1{,}4840 + 0{,}6483 \times 0{,}7170 + 0{,}2125 \times 0{,}7160) =$	19,3394
$\lambda_{TELE} = 23{,}4937 \times (0{,}6483 \times 0{,}1000 + 0{,}2125 \times 0{,}1000) =$	2,0223
$\lambda_{FTK} = 23{,}4937 \times (0{,}6483 \times 0{,}4000 + 0{,}2125 \times 0{,}4000) =$	8,0892
$\lambda_{CRG} = 23{,}4937 \times (0{,}0003 \times 0{,}0385 + 0{,}0110 \times 0{,}0304) =$	0,0082
$\lambda_{SVM} = 23{,}4937 \times (0{,}0003 \times 0{,}9616) =$	0,0076
$\lambda_{LAB} = 23{,}4937 \times (0{,}0110 \times 0{,}2930 + 0{,}1278 \times 0{,}2930 + 0{,}6483 \times 0{,}1465 + 0{,}2125 \times 0{,}1465) =$	3,9187
$\lambda_{RAD} = 23{,}4937 \times (0{,}0110 \times 0{,}2310 + 0{,}1278 \times 0{,}2310 + 0{,}6483 \times 0{,}1155 + 0{,}2125 \times 0{,}1155) =$	3,0895
$\lambda_{TRT} = 23{,}4937 \times (0{,}1278 \times 0{,}1570 + 0{,}6483 \times 0{,}1500 + 0{,}2125 \times 0{,}1500) =$	3,5050
$\lambda_{UDC} = 23{,}4937 \times (0{,}0110 \times 0{,}9696 + 0{,}1278 \times 0{,}2000) =$	0,8523
$\lambda_{ALT} = 23{,}4937 \times (0{,}1278 \times 0{,}5000 + 0{,}6483 \times 1{,}0000 + 0{,}2125 \times 1{,}0000) =$	21,7246
$\lambda_{INT} = 23{,}4937 \times (0{,}0003 \times 1{,}0000 + 0{,}0110 \times 1{,}0000 + 0{,}1278 \times 0{,}4000) =$	1,4688

Tabela 7 – Fluxo de paciente em cada um dos 12 nós (*i*)

e) Estabelecer a qualidade do serviço

Qualquer pronto-socorro almeja fornecer um serviço de alta qualidade, mantendo a

relação custo-benefício no que diz respeito ao uso dos espaços e dos recursos humanos. Geralmente, a liderança em um pronto-socorro conhece bem o volume em que sua qualidade de serviço diminui frente a capacidade disponível, mas não consegue expressar sua compreensão em termos de filas. Os locais de cuidado devem estar cerca de 20% vazios (utilização ≤ 80%), em média, para garantir tempos de espera razoáveis e baixa probabilidade de superlotação e, portanto, minimizar o bloqueio entre as áreas. Sendo assim, especificar a qualidade do serviço em termos de tempos de espera do paciente (W_q^{meta}), taxa de utilização (ρ^{meta}) e probabilidades de superlotação ou *overflow* da área (p_C^{meta}) é mais intuitivo para os gestores deste tipo de sistema. A Tabela 8 descreve a qualidade requerida do serviço.

A utilização é, portanto, apresentada como uma consequência da qualidade do serviço e não como uma variável de projeto. As medidas de desempenho da fila de espera e superlotação explicam melhor a experiência do paciente e sensibilizam os gestores de pronto-socorro. Por exemplo, baixas probabilidades de superlotação são mais importantes em algumas áreas; é mais importante acomodar todos os pacientes nos devidos locais de cuidado do que em uma área de espera. Pacientes aguardando por resultados podem esperar em qualquer espaço aberto no pronto-socorro, enquanto aqueles em assistência requerem mais supervisão e cuidados médicos.

Como cada iteração nesse processo de decisão requer execuções do computador, a vantagem de um modelo baseado em filas (tempo de análise em segundos) e modelos baseados em simulação (tempo de análise em minutos) é significativa. Além disso, não há necessidade de estabelecer condições iniciais, compensar o viés de autocorrelação ou executar várias instâncias para gerar intervalos de confiança.

Área_i	P_{C_i} (%)	w_{q_i} (min.)	ρ^{meta}
Triagem, *TRG*		≤ 05	
Consulta Médica, *DOC*		≤ 15	0,70 < ρ ≤ 0,80
Teleconsulta, *TELE*		≤ 05	
Fast-Track, *FTK*		≤ 05	0,70 < ρ ≤ 0,80
Centro Cirúrgico, *CRG*	< 2		
Sala Vermelha, *SVM*	< 5	= 00	
Laboratório, *LAB*		≤ 15	
Radiologia, *RAD*		≤ 15	
Tratamento, *TRT*			
Unid. Decisão Clínica, *UDC*		≤ 15	0,70 < ρ ≤ 0,80
Alta Médica, *ALT*		≤ 15	0,70 < ρ ≤ 0,80
Unidades Internação, *INT*	< 5	≤ 60	0,70 < ρ ≤ 0,80

Tabela 8 – Qualidade requerida do serviço (fonte: Autor)

f) Prover capacidade aos nós e comparar com as metas

Para uma dada taxa de utilização "*meta*" de uma determinada iteração, cada nó é capacitado e as respectivas medidas de desempenho calculadas. A taxa de utilização dos recursos (ρ) é definido como $\rho = \lambda/c\mu$. Onde λ é a taxa de chegada (pacientes/hora), μ a taxa de serviço (pacientes/hora) e c a quantidade de recursos: servidores (profissionais) ou leitos (locais de cuidado), dependendo da área em análise. O número de servidores (c) é calculado utilizando a Equação (4):

$$c = \frac{\lambda}{\mu \times \rho^{meta}} \tag{4}$$

Foi utilizado a aproximação de Allen-Cunneen para o modelo *M/G/c*, Equação (5), para calcular o tempo médio de espera na fila W_q:

$$\widehat{W_q} = \frac{(\rho \times c)^c p_0}{c \times c!\,(1-\rho)^2}\left(\frac{cv_a^2 + cv_s^2}{2}\right)\frac{1}{\mu} \tag{5}$$

Em que cv_a e cv_s são os coeficientes de variação dos processos de chegada e de serviço, p_0 é a probabilidade da fila estar vazia para um *M/M/c* e $1/\mu$ é o tempo médio de atendimento. A p_0 para uma fila *M/M/c* foi calculada, através da Equação (6), como:

$$P_0 = \left(\sum_{n=0}^{c-1}\frac{(\rho \times c)^n}{n!} + \frac{(\rho \times c)^c}{c!\,(1-\rho)}\right)^{-1} \tag{6}$$

Para estimar a probabilidade de superlotação P_C, contamos com a fila *M/G/c/c*. Observe que esse resultado acomoda distribuições de tempo de serviço que não são exponenciais e é calculado através da Equação (7).

$$P_c = \frac{\frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}}{\frac{\sum_{i=0}^{c}\left(\frac{\lambda}{\mu}\right)^i}{i!}} \tag{7}$$

Dado as premissas de qualidade requerida para cada uma das áreas (nós) do serviço de urgência, conforme Tabela 8, e o volume de pacientes, através das Equações (4), (5), (6) e (7) foram calculados: a necessidade de servidores ou locais de cuidado ($c^{projetado}$), conforme o caso,

a probabilidade de superlotação (P_c), o tempo de espera (W_q) e a utilização (ρ), apresentados na Tabela 9.

Área_i	$c^{instalado}$	$c^{projetado}$	P_{C_i}(%)	w_{q_i}(min.)	ρ
Triagem, *TRG*	2	2	32,3	03,96	0,78
Registro, *REG (nó virtual)*	3	3	-	03,57	0,78
Consulta Médica, *DOC*	12	8	13,5	05,71	0,78
Teleconsulta, *TELE*	0	1	28,8	05,09	0,40
Fast-Track, *FTK*	0	4	12,9	01,82	0,58
Centro Cirúrgico, *CRG*	1	1	01,6	02,00	0,02
Sala Vermelha, *SVM*	2	2	00,0	00,01	0,01
Laboratório, *LAB*	-	-	00,8	03,21	0,52
Radiologia, *RAD*	2	2	14,4	01,64	0,39
Tratamento, *TRT*	30	14	00,0	00,01	0,25
Unid. Decisão Clínica, *UDC*	0	8	07,7	24,54	0,57
Alta Médica, *ALT*	0	6	25,0	47,00	0,97
Unidades Internação, INT	207	207	30,4	70,24	0,73

Tabela 9 – Capacitação dos nós versus qualidade do serviço (fonte: Autor)

g) Calcular o tempo porta-médico

A etapa final do método consiste em estimar o tempo porta-médico (*D2D*), intervalo de tempo médio esperado entre a chegada do paciente ao pronto-socorro e o início da consulta médica. Dentro deste intervalo, encontram-se os tempos de classificação de risco e de registro, bem como as esperas entre estes processos.

Na segmentação do fluxo de pacientes do pronto-socorro, há 2 processos comuns aos 5 perfis de pacientes, o de classificação de risco (*TRG*) e o de registro (*REG*), cujos tempos de espera (na fila) e de atendimento (serviço) devem ser somados. A partir daí os fluxos se dividem e devem ser ponderados quanto ao tempo de deslocamento até o médico (*DOC* e *FTK/TELE*) e a espera (na fila) pela consulta. O tempo porta-médico (*D2D*) pode ser calculado pela Equação (8):

$$\begin{aligned} T_{D2D} = {} & \left(W_{q\,(TRG)} + T_{atendimento\,(TRG)}\right) + \left(W_{q\,(REG)} + T_{atendimento\,(REG)}\right) \\ & + \left[\frac{\lambda_{DOC}}{\lambda_{TRG}} \times \left(T_{deslocamento} + W_{q\,(DOC)}\right)\right] \\ & + \left[\frac{\lambda_{TELE}}{\lambda_{TRG}} \times \left(T_{deslocamento} + W_{q\,(TELE)}\right)\right] \\ & + \left[\frac{\lambda_{FTK}}{\lambda_{TRG}} \times \left(T_{deslocamento} + W_{q\,(FTK)}\right)\right] \end{aligned} \tag{8}$$

$$T_{D2D} = (3{,}97 + 4{,}00) + (3{,}57 + 6{,}00) + \left[\frac{19{,}34}{23{,}49} \times (5{,}00 + 5{,}71)\right] + \left[\frac{2{,}02}{23{,}49} \times (5{,}00 + 5{,}09)\right] + \left[\frac{8{,}09}{23{,}49} \times (5{,}00 + 1{,}82)\right]$$

$T_{D2D} \cong 29$ min.

Todos os termos na Equação (8) foram calculados anteriormente, conforme Tabelas 7 (λ) e 9 (W_q), exceto os tempos de deslocamento ($T_{deslocamento}$), que foram estimados em 5 minutos, independente da classificação de risco e do fluxo do paciente.

4.3. Análise dos resultados obtidos

As curvas de *trade-off* desempenham um papel importante e auxiliam na tomada de decisão e no planejamento dos serviços de urgência. O enfoque é no *trade-off* entre o tempo médio de permanência e a capacidade de atendimento aos pacientes. Uma alocação insuficiente de capacidade nesses serviços pode causar tempos de espera elevados e longos *leadtimes*, por outro lado, o excesso de capacidade pode resultar em desperdício de recursos onerosos, devido aos baixos níveis de utilização. Assim, uma questão central nesta dissertação é a seleção entre as várias configurações para a rede ou, mais especificamente, como os recursos, sejam eles humanos ou físicos, devem ser adequadamente distribuídos para prover capacidade nas várias áreas.

Curvas de *trade-off* são curvas de fronteira ótima que indicam as vantagens (e desvantagens) da troca de um ponto por outro da fronteira. No caso de uma curva de *trade-off* entre o tempo médio de permanência (*LOS*) e a capacidade, para cada recurso, a curva indica o ponto com o *leadtime* ou *LOS* mínimo.

A Figura 21 ilustra as alternativas da alta direção do pronto-socorro ao optar por segmentar ou não os pacientes. Ao não segmentar os pacientes, o protocolo de classificação é utilizado apenas como critério de *priorização* a fim de direcioná-los aos 12 consultórios disponíveis (sem *fast-track/teleconsulta*). Já ao segmentar utilizando o mesmo protocolo é criado um fluxo específico para os pacientes de baixo risco clínico, que representam 85% do total. Neste caso, a ideia é criar 2 caminhos: (i) um fluxo normal (*DOC*) e (ii) um fluxo rápido (*FTK/TELE*), alocando os profissionais médicos de acordo com as demandas. Lembrando que os pacientes candidatos ao *fast-track* ou à teleconsulta devem demandar apenas uma consulta

médica ou no máximo um exame (pactuado com a equipe do SADT) cujo tempo de resposta (*TAT*) seja inferior a 30 minutos. Para todos os casos com *fast-track*, foi mantido sempre 1 consultório de teleconsulta, com 10% do fluxo (2 pacientes/hora), afim de acomodar todos os profissionais médicos.

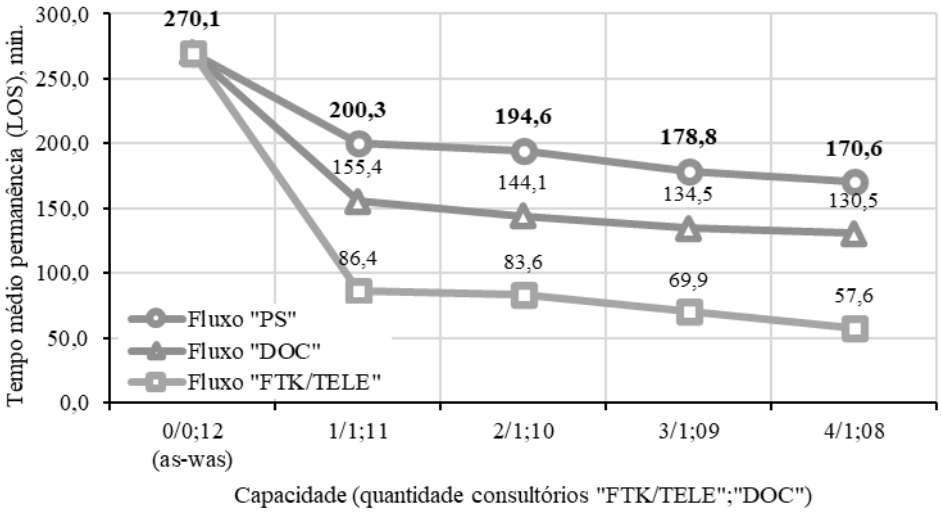


Figura 21 – Curva de *trade-off* entre a capacidade e o tempo médio de permanência (*LOS*)

Os pontos da Figura 21 representam a correlação entre o tempo médio de permanência (*LOS*) e a combinação de consultórios *FTK/TELE* e *DOC*. Por exemplo, o ponto (0/0;12) representa 0 consultórios *fast-track*/teleconsulta e 12 consultórios normais. Dentre todas as combinações testadas, a que apresenta menor *LOS* com 170,6 minutos é o ponto (4/1;08) com 4 consultórios *fast-track*, 1 consultório teleconsulta (*FTK/TELE*) e 8 consultórios normais (*DOC*). Neste mesmo ponto, o fluxo normal "DOC" de pacientes apresenta um *LOS* de 130,5 minutos e o fluxo rápido *"FTK/TELE"* um *LOS* de 57,6 minutos.

Já, na Figura 22, observa-se a curva de *trade-off* entre a capacidade alocada nos consultórios de *fast-track*/teleconsulta e o tempo porta-médico (*D2D*). O ponto (1/1;11) representa 1 consultórios *fast-track*, 1 consultório teleconsulta e 11 consultórios normais, o maior tempo porta-médico testado, com 38,4 minutos. Dentre todas as configurações testadas, a que apresenta menor tempo porta-médico com 29,1 minutos é o ponto (4/1;08) com 4 consultórios *fast-track*, 1 consultório teleconsulta e 8 consultórios normais. Neste mesmo ponto, o fluxo normal "DOC" de pacientes apresenta um *D2D* de 26,4 minutos e o fluxo rápido *"FTK/TELE"* um *D2D* de 20,3 minutos.

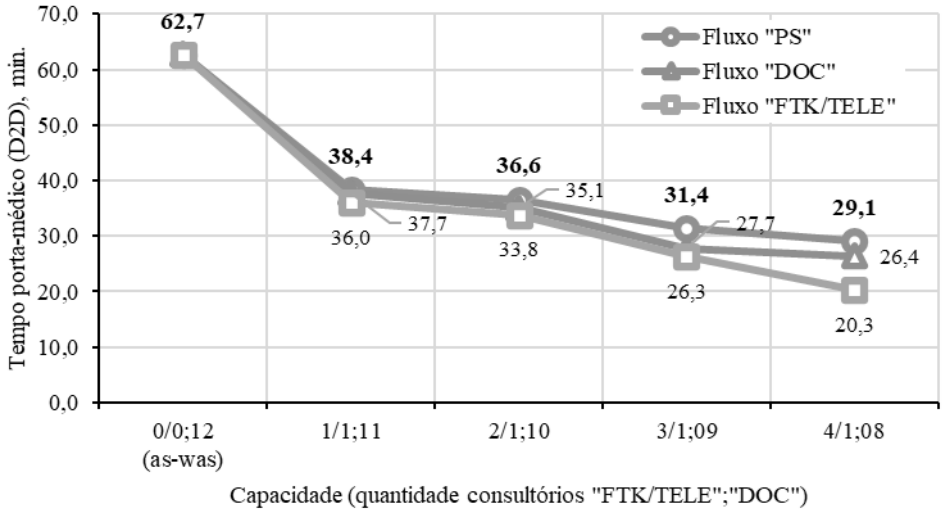


Figura 22 – Curva de *trade-off* entre a capacidade e o tempo porta-médico (*D2D*)

A Figura 23 traz o comparativo entre o fluxo rápido (*FTK/TELE*) e o fluxo normal (*DOC*) isoladamente testados e, a combinação deles (*PS*), na melhor configuração obtida, para o tempo médio de permanência (*LOS*) e os três intervalos de tempo (porta-médico, médico-decisão e decisão-saída) do pronto-socorro. Como pode-se observar, a contribuição da segmentação dos pacientes para redução dos tempos de passagem, com a adoção do fluxo rápido (*FTK/TELE*), é bastante expressiva.

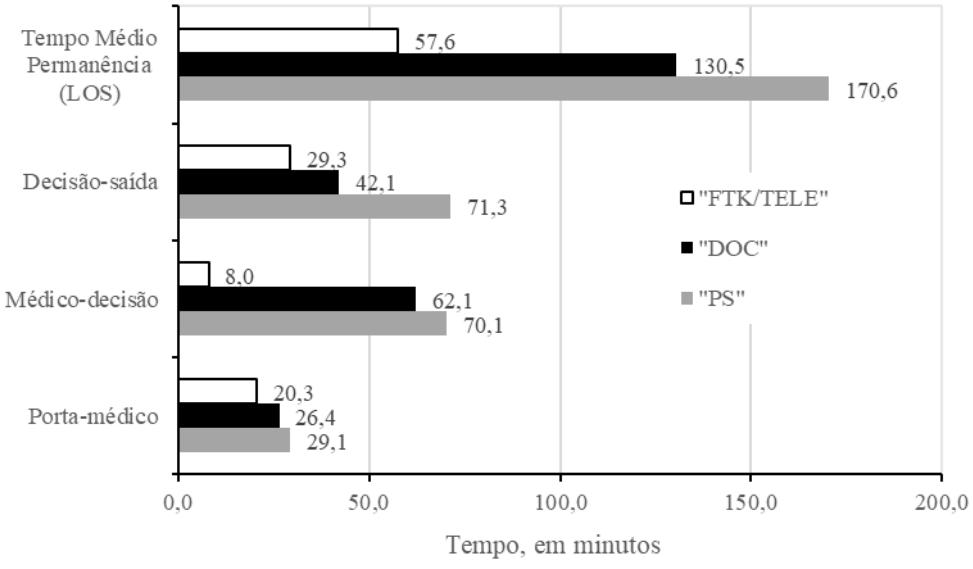


Figura 23 – Comparativo de tempos de pacientes no pronto-socorro

A Figura 24 traz a curva de *trade-off* entre o tempo médio de permanência (*LOS*) e o volume anual de pacientes. O ponto (4/1;08) apresenta menor tempo de permanência com 170,6 minutos e volume pouco superior a 180 mil pacientes por ano, alcançado com 4 consultórios

fast-track, 1 teleconsulta e 8 normais. O ponto (*as-was*) refere-se à situação anterior à mudança, com 116,4 mil pacientes por ano e, quando comparado com o ponto (4/1;08), estima-se um aumento anual de 64,5 mil pacientes, o que representa um acréscimo de 55%. Considerando um ticket médio de R$ 250 por paciente, projeta-se, portanto, um incremento de receita, superior a R$ 15 milhões por ano.

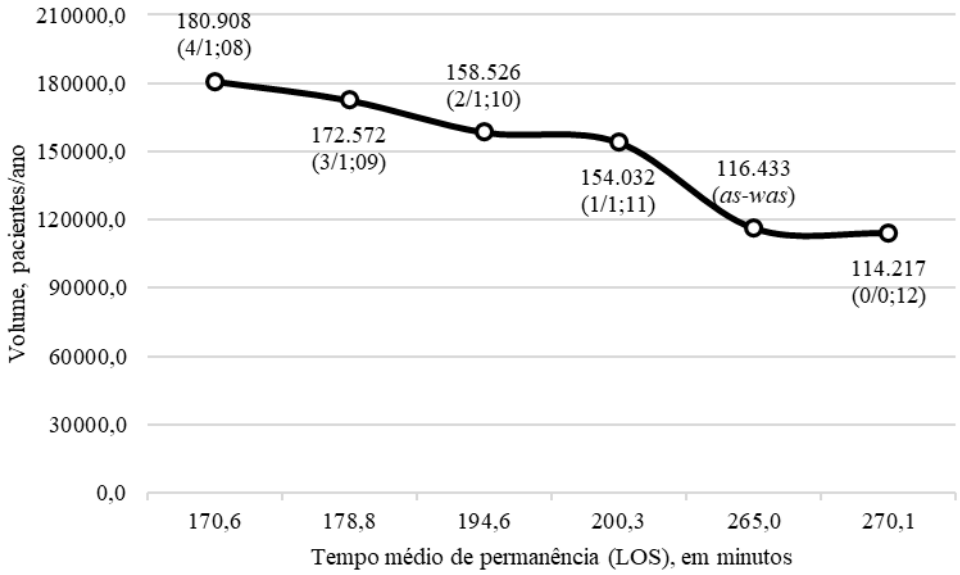


Figura 24 – Curva de *trade-off* entre o tempo médio de permanência (LOS) e o volume anual

A Figura 25 ilustra a curva de *trade-off* entre a taxa média de chegada de pacientes nos 4 consultórios *fast-track*, de acordo com o percentual do fluxo desviado, e o percentual de pacientes que visitam o pronto-socorro e realizam exames de laboratório e/ou raio-x. O impacto do *fast-track* na redução do número de exames realizados é significativo. Com um menu de exames bem estruturado através de protocolos clínicos e a garantia da priorização destes para os pacientes do *fast-track*, o pronto-socorro pode economizar recursos financeiros e materiais, além de liberar equipamentos e profissionais para atender a um maior número de pacientes. Isso resulta em uma maior eficiência operacional e uma melhor utilização dos recursos disponíveis.

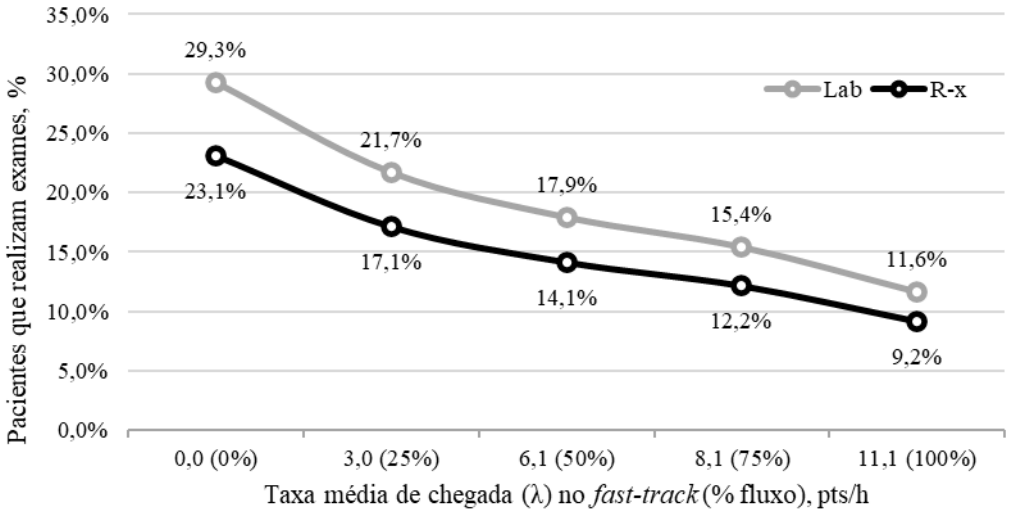


Figura 25 – Curva de *trade-off* entre taxa de chegada e pacientes que realizam exames

Vale lembrar que os profissionais médicos que atuam no consultório *fast-track* precisam ser experientes e estar habituados a esse perfil de atendimento.

4.4. Análise das unidades de internação

É comum, pacientes aguardarem pela internação no pronto-socorro (tempo de *boarding*), ocupando um local de cuidado que se destinaria a um novo paciente e superlotando, muitas vezes, as instalações. Tal situação se deve à dificuldade do hospital em girar seus leitos nas unidades de internação. Não é o caso do hospital "H" como se pode comprovar, a seguir, através do cálculo da utilização dos leitos nas suas unidades de internação.

Levando a variável μ do denominador para o numerador, a taxa de serviço (pacientes/dia) passa a ser tempo médio de permanência (dias/paciente) e, portanto, tem-se a seguinte equação (9):

$$\rho = \frac{\lambda\,(\frac{\text{pts}}{\text{dia}}) \times \mu^{-1}(\frac{\text{dias}}{\text{pts}})}{\text{c}\ (\#\text{leitos})} \tag{9}$$

Dada a taxa de conversão de 6,25%, ou seja, do total de pacientes que visitam o pronto-socorro a parcela cuja decisão médica é pela "internação" representa 36 pacientes/dia (#internações/dia), portanto, substituindo os valores na equação (9) tem-se:

$$\rho = \frac{\#\text{internações/dia} \times \text{tempo médio de permanência}}{\#\text{leitos}} = \frac{36 \times 2{,}70}{207} = 0{,}47$$

Uma subutilização do recurso leito, ao considerar $0{,}60 < \rho \leq 0{,}80$ como a faixa adequada de utilização. Além disso, considerando uma utilização dos 207 leitos (*c=207*) nas unidades de internação de 80% (*ρ=0,8*), o hospital "H" teria um potencial de internação de 61 pacientes/dia (atual 36 pacientes/dia), como demonstrado no cálculo abaixo (ponto (2,70; 61) da curva ρ=0,80 na Figura 26):

$$\#\text{internações} = \frac{\text{utilização } (\rho) \times \#\text{leitos}}{\text{tempo médio de permanência}} = \frac{0{,}80 \times 207}{2{,}70} \cong 61$$

A Figura 26 apresenta a curva de *trade-off* entre o tempo médio de internação (*LOH*) e o número de internações por dia, ou seja, mesmo levando o tempo médio de internação para 3,0 dias ainda existiria capacidade para internar 55 pacientes, conforme ponto (3,00; 55) da curva ρ=0,80, superior aos atuais 36 pacientes.

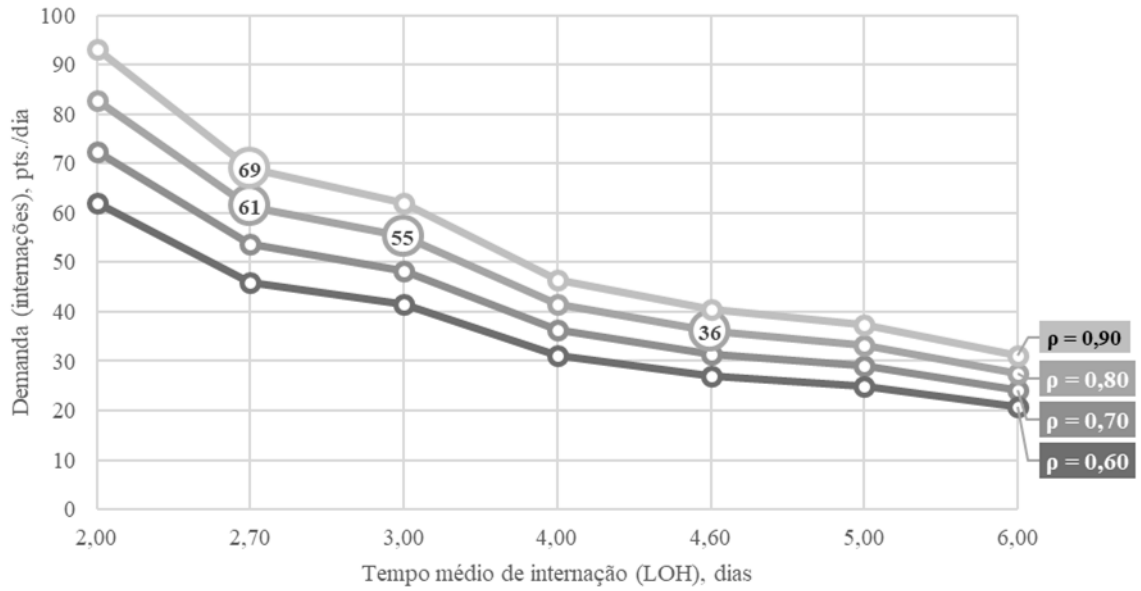


Figura 26 – Curva de *trade-off* entre o tempo médio de internação e o #internaçãos/dia (*ρ*=0,8)

Mais ainda, mantido o tempo médio de internação em 2,7 dias, conforme representado pelo ponto (2,70; 69) da curva ρ=0,90 na Figura 26, se houver um aumento expressivo de pacientes em *boarding* no pronto-socorro e o hospital decidir por operar com uma utilização dos leitos na ordem de 90%, poder-se-á internar 69 novos pacientes/dia.

Agora, observando a mesma curva de *trade-off* representada pela Figura 26, mantido o número de internações em 36 pacientes por dia, o tempo médio de internação, considerando uma utilização média dos leitos de 80%, poderia ser de 4,6 dias, conforme ponto (4,60; 36) da curva ρ=0,80.

$$\text{Tempo médio de permanência } = \frac{0{,}80 \times 207 \text{ leitos}}{36 \text{ pacientes/dia}} = 4{,}60 \text{ dias}$$

Ao analisar as curvas de *trade-off*, os gestores hospitalares podem identificar onde ocorrem os melhores *trade-offs* entre eficiência, custos e qualidade. Isso ajuda a otimizar os processos e melhorar o desempenho geral da organização.

5. CONCLUSÕES E TRABALHOS FUTUROS

A superlotação de serviços de urgência representa grave problema do sistema de saúde e a demanda espontânea dos pacientes com doenças simples amplifica exageradamente o número de atendimento nessas unidades. Entretanto, a assistência médica pode ocorrer em condições precárias, quer seja em relação à demora no atendimento, qualidade do serviço prestado e acomodação no ambiente de espera, no aguardo da avaliação, diagnóstico e tratamento desses pacientes no ambiente do pronto-socorro, destinado ao atendimento de casos graves e complexos. As "portas abertas", garantindo o desejado atendimento médico, incorrem na ausência de restrições ao acesso, exigem ampla complacência do ambiente e da equipe multidisciplinar, sob risco de queda na qualidade do serviço prestado. Na rotina diária pode-se mensurar tempo de espera prolongado, superlotação das salas de espera, necessidade de leitos extra, em ambiente desprovido da necessária e desejada privacidade e conforto, no aguardo de vagas nas unidades de internação ou terapia intensiva. Em uma análise geral dos atendimentos realizados, constata-se que mais de 60% dos diagnósticos não estão relacionados a situações de urgência e emergência. Nesse enfoque, podemos constatar que a superlotação nos equipamentos de saúde, destinados ao atendimento de urgências, pode representar dificuldade no atendimento médico nas unidades de saúde de menor complexidade. Amplificando essa demanda espontânea, acrescente-se o reconhecido menor poder de resolutividade das Unidades de Pronto Atendimento (*UPA*) na periferia, aliado ao notório poder de atração dos hospitais, habitualmente de "portas abertas", incrementando a migração dos clientes para o pronto-socorro das instituições hospitalares.

Nessas condições, o excesso de demanda de casos simples pode prejudicar o atendimento médico especializado, congestionando as equipes diante de doenças sem maior gravidade, com eventual detrimento aos pacientes em real situação de emergência e risco de morte, carentes de cuidados intensivos. O desgaste das equipes e do ambiente contribui para o descontentamento de pacientes e de toda a equipe de trabalho, proporcionando condições que promovem o estresse e consequentes falhas no atendimento ao paciente. Dentro dessa rotina diária potencialmente caótica, os pacientes com doenças de menor complexidade coabitam e disputam a desejada e merecida atenção, deslocando involuntariamente os recursos destinados àqueles realmente em estado grave. Diante dessas condições, urge planejar rotinas de trabalho e implementar projetos e propostas destinadas a garantir o atendimento eficiente, humanizado e digno, atenuando a pressão dessa excessiva demanda.

A metodologia aqui proposta permite modelar qualquer combinação de pacientes (riscos

clínicos), volume de chegada e desempenho operacional, para um novo modelo de segmentação do fluxo de pacientes no pronto-socorro. Cada uma das áreas modeladas pode ser dimensionada de acordo com as métricas de desempenho almejadas. As variações diárias (picos) e mensais (sazonalidades) podem e devem ser consideradas. Variações nos volumes de atendimento podem ser simulados e as capacidades necessárias facilmente localizadas. Melhora significativa no tempo de resposta para análise de cenários e facilidade de uso quando comparado com simulação por computador. A metodologia pode ser estendida para todo o hospital, ou pelo menos para aquelas áreas consideradas chave, como por exemplo, o centro cirúrgico e as unidades de internação (UTIs e enfermarias).

O pronto-socorro "H" agora se destaca como um modelo de agilidade e eficácia. A aplicação da metodologia resultou em uma redução significativa nos tempos de espera, permitindo uma resposta mais rápida e eficiente às emergências médicas. A eliminação de desperdícios e a otimização no fluxo dos pacientes não apenas aumentaram a eficiência operacional, mas também melhoraram substancialmente a experiência do paciente. A equipe assistencial, agora, melhor estruturada, consegue oferecer um atendimento de qualidade superior. A transformação não apenas aprimorou os indicadores de desempenho, conforme destacado na Tabela 10, mas também redefiniu o padrão de excelência, proporcionando um ambiente mais seguro e eficiente para cuidados emergenciais de saúde.

Indicadores	***As-was***	***As-is***	**Resultado**
Tempo médio de permanência (*LOS*), min.	265	171	-36%
Tempo porta-médico (*D2D*), min.	65	29	-55%
Tempo médico-decisão, min.	140	70	-50%
Tempo de *boarding*, min.	60	≤ 45	-25%
Taxa de evasão (*LWBS*), %	2,2	< 1,4	-36%
Índice de satisfação do cliente (*NPS*)	37	≥ 75	+103%
Volumetria, kpts./ano	116,4	180,9*	+55%

Tabela 10 – Resultado alcançado com implantação da metodologia (*potencial)

O *benchmarking* é uma das ferramentas mais importantes para sobrevivência e crescimento das organizações. Com ele, atividades bem-sucedidas são tomadas como referência, e suas práticas viram padrões a serem seguidos, por quem deseja evoluir em um segmento comparável. Para isso, as metas deixam de ser pautadas em ideias e imaginação. O alvo estará embasado naquilo que outras organizações já concretizaram e que pode ser usado como referência por outras instituições, com objetivos realistas.

Outro aspecto, que deve ser considerado em pesquisas futuras, é a estrutura dinâmica e

híbrida dos sistemas de saúde dos tempos atuais, o que enfatiza a importância de combinar os modelos matemáticos e gerenciais com os aspectos tecnológicos. Isto pode abrir a porta para novos estudos de campo e métodos completamente inovadores. A pandemia de Covid-19 sublinhou a importância de uma resposta rápida e dinâmica nos prontos-socorros. Mesmo em um momento de "não crise", com crescimento da demanda, surge a necessidade de uma resposta rápida, em um momento de elevada incerteza e de orçamento limitado, enfatizando a importância da implementação eficaz de métodos de filas; mesmo nessas condições, espera-se que a técnica ora proposta, bem como as futuras, aumente a eficiência e auxiliem as equipes médicas na prestação de cuidados adequados aos pacientes do serviço de urgência.

6. REFERÊNCIAS

1. Allen, A.O. (1978). *Probability, statistics, and queueing theory with computer science applications*. Section 5.5: approximations. New York: Academic Press; 1978 p. 217–25.
2. Allen, A.O. (1990). *Probability, Statistics and Queueing Theory: with computer science applications*. 2nd edition. Academic Press. Boston.
3. Allon, G.; Deo, S.; Lin, W. (2013). *The impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence*. Operations research, 61(3), 544-562.
4. Almehdawe, E.; Jewkes, B.; He, Q.M. (2013). *A Markovian queueing model for ambulance offload delays*. European Journal of Operational Research, 226(3), 602-614.
5. Apellaniz, J.S. *et al.* (2023). *Leveraging Telemedicine to Reduce ED Overcrowding: The Quironsalud Virtual Urgent Care Program*, NEJM Catalyst Innovation in Care Delivery Vol. 4 No. 8, DOI: 10.1056/CAT.22.0422.
6. Armony, M.; Israelit, S.; Mandelbaum, A.; Marmor, Y.N.; Tseytlin, Y.; Yom-Tov, G.B. (2015). *Patient Flow in Hospitals: A Data-Based Queueing-Science Perspective*. Working Paper, New York University, New York.
7. Armony, M.; Mandelbaum, A. (2011). *Routing and staffing in large-scale service systems: The case of homogeneous impatient customersand heterogeneous servers*. Oper. Res., 59(1), 50–65.
8. Au, L.; Byrnes, G.B.; Bain, C. A.; Fackrell, M.; Brand, C.; Campbell, D.A.; Taylor, P.G. (2009). *Predicting overflow in an emergency department*. IMA J. Manage. Math., 20, 39–49.
9. Au-Yeung, S.W.M.; Harrison, P.G.; Knottenbelt, W.J. (2006). *A queueing network model of patient flow in an accident and emergency department*. Proceedings of the 20th Annual European and Simulation Modelling Conference. 60–67.
10. Au-Yeung, S.W.M.; Harrison, P.G.; Knottenbelt, W.J. (2007). *Approximate queueing network analysis of patient treatment times*. Proceedings of the 2nd International Conference on Performance Evaluation Methodologies and Tools. 1–12.
11. Batt, R.J.; Terwiesch, C. (2013). *Waiting patiently: An empirical study of queue abandonment in an emergency department*. Working Paper, University of Pennsylvania, Philadelphia, PA.
12. Bekker, R.; de Bruin, A.M. (2009). *Time-dependent analysis for refused admissions in clinical wards*. Ann. Oper. Res., 178(1), 45–65.
13. Bitran, G.R.; Morabito, R. (1995). Manufacturing system design: tradeoff curve analysis. Production and Operations Management Society. https://doi.org/10.1590/S0104-530X1996000200001
14. Bitran, G.R.; Morabito, R. (1999). *AN OVERVIEW OF TRADEOFF CURVES IN MANUFACTURING SYSTEMS DESIGN*. Production and Operations Management, 8(1), 56-75. https://doi.org/10.1111/j.1937-5956.1999.tb00061.x
15. Bitran, G.R.; Sarkar, D. (1994). *Targeting Problems in Manufacturing Queueing Networks - An Iterative Scheme and Convergence*. European Journal of Operational Research, Vol. 76, No. 3, 501-510.
16. BRASIL. Ministério da Saúde. Secretaria de Ciência, Tecnologia, Inovação e Insumos Estratégicos em Saúde. Departamento de Ciência e Tecnologia. Síntese de evidências para políticas de saúde: congestão e superlotação dos serviços hospitalares de urgências. Brasília: Ministério da Saúde; EVIPNet Brasil, 2020. 81 p.
17. Bretthauer, K.M.; Heese, S.; Pun, H.; Coe, E. (2011). *Blocking in healthcare operations: A new heuristic and an application*. Prod. Oper. Manage., 20(3), 375–391.

18. Brockmeyer, E.; Halstrom e Jensen, A. (1948). *The life and works of A. K. Erlang*. Transactions of the Danish Academy of Technical Sciences, No. 2, pp. 1-288.
19. Broyles, J.R.; Cochran, J.K. (2007). *Estimating business loss to a hospital emergency department from patient reneging by queuing-based regression*. Proceedings of the 2007 Industrial Engineering ResearchConference, 613–618.
20. Broyles, J.R.; Cochran, J.K. (2011). *A queuing-base statistical approximation of hospital emergency department boarding*. In Proceedings of the International Conference on Computers and Industrial Engineering.
21. Calabrese, J.M. (1992). *Optimal Workload Allocation in Open Networks of Multiserver Queues*. *Management Science*, 38, 12, 1792–1802.
22. Chowdhury, N.M.; Riddles, L.; MacKenzie R.S. (2018). *Using Queuing Theory to Reduce Wait, Stay in Emergency Department*. American Association for Physician Leadership.
23. Cochran, J.K.; Bharti, A. (2006). *Stochastic bed balancing of na obstetrics hospital*. Health Care Manage. Sci., 9, 31–45.
24. Cochran, J.K.; Broyles, J.R. (2010). *Developing nonlinear queuing regressions to increase emergency department patient safety: Approximating reneging with balking*. Comput. Ind. Eng., 59, 378–386.
25. Cochran, J.K.; Roche, K.T. (2009). *A multi-class queuing network analysis methodology for improving hospital emergency department performance*. Comput. Oper. Res., 36(5), 1497–1512.
26. Dark, C.; Canellas, M.; Mangira, C.; Jouriles, N.; Simon E.L. (2020). *Estimates of throughput and utilization at freestanding compared to low-volume hospital-based emergency departments*. JACEP Open. 2020; 1: 1297–1303.
27. de Bruin, A.M.; van Rossum, A.C.; Visser, M.C.; Koole, G.M. (2007). *Modeling the emergency cardiac inpatient flow: an application of queuing theory*. Health Care Manage. Sci., 10,125–137.
28. de Vericourt, F.; Jennings, O.B. (2008). *Nurse-to-Patient Ratios in Hospital Staffing: A Queueing Perspective*. ESMT Working Paper No. 08-005, Available at SSRN: https://ssrn.com/abstract=1162729 or http://dx.doi.org/10.2139/ssrn.1162729.
29. de Vèricourt, F.; Jennings, O.B. (2011). *Nurse staffing in medical units: A queueing perspective*. Operations Research, 59(6), 1320–1331.
30. Derlet, R.W.; Richards, J.R. (2000). *Overcrowding in the nation's emergency departments: complex causes and disturbing effects*. Annals of emergency medicine, 35(1), 63-68.
31. Derlet, R.W.; Richards, J.R. (2002). *Emergency department overcrowding in florida, new york, and texas*. Southern medical journal, 95(8), 846-850.
32. Ding, Y.; Park, E.; Nagarajan, M.; Grafstein, E. (2019). *Patient prioritization in emergency department triage systems: an empirical study of Canadian Triage and Acuity Scale (CTAS)*. Manuf Serv Oper Manag 21(4):723–741. https://doi-org.ez27.periodicos.capes.gov.br/10.1287/msom.2018.0719
33. Disney, R.L.; Konig, D. (1985). *Queueing Networks: A Survey of Their Random Processes*. SIAM Review, Vol. 27, 335-403.
34. Elalouf, A.; Wachtel, G. (2016). *An alternative scheduling approach for improving emergency department performance*. Int J Prod Econ 178:65–71. https://doi-org.ez27.periodicos.capes.gov.br/10.1016/j.ijpe.2016.05.002
35. Elalouf, A.; Wachtel, G. (2017). *Using the "Floating Patients" method to balance crowding between the hospital emergency department and other departments*. Comput Ind Eng 110:289–296. https://doi-org.ez27.periodicos.capes.gov.br/10.1016/j.cie.2017.06.023

36. Erlang, A.K. (1917). *Solution of Some Problems in the Theory of Probabilities of Some Significance in Automatic Telephone Exchanges*. Post Office Electrical Engineer's Journal, 10, 189–197.
37. Filippatos, G.; Evridiki, K. (2015). *The effect of Emergency department crowding on patient outcomes*. Health Science Journal 2015; 9(16): 1–6.
38. Fitzgerald, K.; Pelletier; L.; Reznek, M.A. (2017). *A queue-based Monte Carlo analysis to support decision making for implementation of an emergency department fast track*. J Healthcare Eng 2017:1–8.https://doi-org.ez27.periodicos.capes.gov.br/10.1155/2017/6536523
39. FOGLIATTI, Maria Cristina; MATTOS, Néli Maria Costa (2007). *Teoria de filas*. Rio de Janeiro: Interciência, 1–290.
40. Folake, A.O.; Agu, M.N.; Okebanama, U.F. (2020). A*pplication of Queue Model in Health Care Sector*. Int Res J Adv Eng Sci. 5(3):48–50. Available online at http://irjaes.com/wp-content/uploads/2021/01/IRJAES-V5N2P274Y20.pdf
41. Fomundam, S.; Herrmann, J. (2007). *A survey of Queuing Theory Applications in healthcare*. ISR Technical Report 24.
42. Gallivan, S.; Utley, M.; Treasure, T.; Valencia, O. (2002). *Booked inpatient admissions and hospital capacity: mathematical modelling study*. Brit. Med. J., 324, 280–282.
43. Gershwin, S.B. (1994). *Manufacturing Systems Engineering*. Prentice Hall, Englewood Cliffs, NJ.
44. Ghani, N.A. (2012). *Multi-server queuing maximum availability location problem with stochastic travel times*. Proceedings of the World Congress on Engineering, 1, 137–143.
45. Goodacre, S.; Webster, A. (2005). *Who waits longest in the emergency department and who leaves without being seen?*. Emerg. Med. Journal; 22(2):93-6. doi: 10.1136/emj.2003.007690. PMID: 15662055; PMCID: PMC1726682.
46. Gordon, R.; Graber, M. and Franklin, N. (2002). *Reducing diagnostic errors in medicine: what's the goal?* Academic Medicine, 77(10), 981-992.
47. Green, L.V. (2002). *How many hospital beds?*. INQUIRY: The Journal of Health Care Organization, Provision, and Financing, 39(4), 400-412.
48. Green, L.V.; Kolesar, P.J.; Soares, J. (2001). *Improving the SIPP approach for staffing service systems that have cyclic demands*. Oper.Res., 49(4), 549–564.
49. Green, L.V.; Kolesar, P.J.; Whitt, W. (2007). *Coping with time-varying demand when setting staffing requirements for a servisse system*. Prod. Oper. Manage., 16(1), 13–39.
50. Green, L.V.; Nguyen, V. (2001). *Strategies for cutting hospital beds: The impact on patient service*. Health Serv. Res., 36(2), 421–442.
51. Green, L.V.; Soares, J.; Giglio, J.F.; Green, R.A. (2006). U*sing queuing theory to increase the effectiveness of emergency department provider staffing*. Academic Emergency Medicine;13(1):61–68.
52. Gross, D.; Harris, C.M. (1998). *Fundamentals of queueing theory*. 3rd ed. Section 4.2: Open Jackson networks. New York: Wiley; p. 174–83.
53. Haussmann, R.K. (1970). *Waiting time as an index of quality of nursing care*. Health Services Research 5: 92-105.
54. HealthTech Briefing Report (2006). *Key trends in emergency and trauma services*. Health Technology Center, http://www.healthtechcenter.org/.
55. Hijry, H.; Olawoyin, R. (2022). *Predicting Patient Waiting Time in the Queue System Using Deep Learning Algorithms in the Emergency Room*. International Journal of Industrial Engineering and Operations Management 3(1):33 – 45 DOI:10.46254/j.ieom.20210103.
56. Hillier, F.; Lieberman, G.S. (1990). *Introduction to Operations Research*. Mc GrawHill, 1990. https://doi-org.ez27.periodicos.capes.gov.br/10.1002/emp2.12318

57. Hsu, L.F.; Tapiero, C.S.; Lin, C. (1993), "Network of Queues Modeling in Flexible Manufacturing Systems: A Survey," *Recherche Operationnelle l'Operations Research*, 27, 2, 201–248.
58. Hu, X.; Barnes, S.; Golden, B. (2018). *Applying queueing theory to the study of emergency department operations: a survey and a discussion of comparable simulation studies*. International Transactions in Operational Research, 25(1), pp. 7-49*.
59. Huang, J. (2013). *Patient flow management in emergency departments*. Ph.D. Dissertation, National University of Singapore.
60. Huang, J.; Carmeli, B.; Mandelbaum, A. (2013). *Control of Patient Flow in Emergency Departments, or Multiclass Queues with Deadlines and Feedback*. Working Paper, National University of Singapore.
61. Izady, N.; Worthington, D. (2012). *Setting staffing requirements for time dependent queueing networks: The case of accident and emergency departments*. European Journal of Operational Research, 219(3), 531-540.
62. Kao, E.P.C.; Tung, G.G. (1981). *Bed allocation in a public healthcare delivery system.* Manage. Sci., 27(5), 507–520.
63. Kim, S.; Horowitz, I.; Young, K.K.; Buckley, T.A. (1999). *Analysis of capacity management of the intensive care unit in a hospital*. Eur. J. Oper. Res.,115, 36–46.
64. Kleinrock, L. (1976). *Queueing Systems, Vol. II: Computer Applications*. John Wiley & Sons, New York.
65. Koenigsberg, E. (1982). *Twenty Five Years Of Cyclic Queues And Closed Queue Networks: A Review*. Journal of the Operational Research Society, Vol. 33, 605-619.
66. Koizumi, N.; Kuno, E.; Smith, T.E. (2005). *Modeling patient flows using a queuing network with blocking*. Health Care Manage. Sci., 8, 49–60.
67. Komashie, A.; Mousavi, A.; Clarkson, P.J.; Young, T. (2015). *An integrated model of patient and staff satisfaction using queuing theory*. IEEE journal of translational engineering in health and medicine, 3, 1-10.
68. Kouvelis, P.; Tirupati, D. (1991). *Approximate Performance Modeling and Decision Making for Manufacturing Systems: A Queueing Network Optimization Framework*. Journal of Intelligent Manufacturing, Vol. 2, 107-134, 1991.
69. Lakshmi, C.; Sivakumar, A.I. (2013). *Application of queueing theory in healthcare: a literature review*. Oper. Res. Healthc. 2. 25-39.
70. Lin, L.; Wang, Q.; Sadek, A.W. (2014). *Border crossing delay prediction using transient multi-server queueing models*. Transportation Research Part A: Policy and Practice, 64, 65-91.
71. Madsen, T. L.; Kofoed-Enevoldsen, A. (2011). *Five easy equations for patient flow through an emergency department*. Danish medical bulletin, 58(10), A4318.
72. Maman, S. (2009). *Uncertainty in the Demand for Service: The Case of Call Centersand Emergency Departments*. Master's thesis, Technion – Israel Institute of Technology.
73. Mandelbaum, A.; Momčilović, P.; Tseytlin, Y. (2012). *On fair routing from emergency departments to hospital wards: QED queues with heterogeneous servers*. Management Science, 58(7), 1273-1291.
74. Marchal, G. (1978). *Some Simpler Bounds on the Mean Queueing Time*, Operations Research, Vol. 26, pp. 1083-1088.
75. Marianov, V.; ReVelle, C. (1996). *The queueing maximal availability location problem: A model for the siting of emergency vehicles*. Eur. J. Oper. Res., 93, 110–120.
76. Mayhew, L.; Smith, D. (2008). *Using queuing theory to analyse the government's 4-h completion time target in accident and emergency departments*. Health care management science, 11(1).

77. Morabito, R. (1998). *Análise de curvas de trade-off baseada em teoria de redes de filas para o projeto e planejamento de sistemas discretos de manufatura*. São Carlos, tese de livre-docência, EESC/USP, 136pg.
78. Osorio, C.; Bierlaire, M. (2009). *An analytic finite capacity queueing network model capturing the propagation of congestion and blocking*. Eur. J. Oper. Res., 9, 996–1007.
79. Ozcan; Y.A. (2005). *Forecasting. In: Quantitative methods in health care management*, California: Josey-Bass; 2005. p. 10–44 [Chapter 2].
80. Palvannan, R.K.; Teow, K.L. (2012). *Queueing for healthcare*. Journal of medical systems, 36, 541-547.
81. Pesquisa Instituto Datafolha (2018). *Opinião dos brasileiros sobre o atendimento público na área de saúde*. Conselho Federal de Medicina (CFM). https://portal.cfm.org.br/images/PDF/datafolha_sus_cfm2018.pdf
82. Qandeel, M.S.; Al-Qudah, I.K.; Nayfeh R. *et al.* (2023). *Analyzing the queuing theory at the emergency department at King Hussein cancer center*. BMC Emerg Med 23, 22. https://doi.org/10.1186/s12873-023-00778-x.
83. Reichheld, F.F. (2003). *The one number you need to grow*. Harvard Business Review, 81(12):46-54, 124. PMID: 14712543.
84. Ridge, J.C.; Jones, S.K.; Nielsen, M.S.; Shahani, A.K. (1998). *Capacity planning for intensive care units*. Eur. J. Oper. Res.,105,346–355.
85. Roche, K.T.; Cochran, J.K. (2007). *Improving patient safety by maximizing fast-track benefits in the emergency department: A queuing network approach*. Proceedings of the 2007 Industrial Engineering Research Conference, 619–624.
86. Rowe, B.; Bond, K.; Ospina, M.; Blitz, S.; Afilalo, M.; Campbell, S.; Schull, M. (2006). *Frequency, determinants, and impact of overcrowding in emergency departments in Canada: a national survey of emergency department directors*. https://www.cadth.ca/sites/default/files/pdf/320c_overcrowding_tr_e_RIB.pdf
87. Saghafian, S.; Hopp, W.J.; Van Oyen, M.P.; Desmond, J.S.; Kronick, S.L. (2012). *Patient streaming as a mechanism for improving responsiveness in emergency departments*. Oper. Res., 60(5), 1080–1097.
88. Saghafian, S.; Hopp, W.J.; Van Oyen, M.P.; Desmond, J.S.; Kronick, S.L. (2014). *Complexity-augmented triage: A tool for improving patient safety and operational efficiency*. Manuf. Serv. Oper. Mang., 16(3), 329–345.
89. Saghafian, S.; Austin, G.; Traub, S.J. (2015). *Operations research/management contributions to emergency department patient flow optimization: Review and research prospects*. IIE Transactions on Healthcare Systems Engineering 2015. 5(2), pp. 101-123*.
90. Sharif, A. B.; Stanford, D.A.; Taylor, P.; Ziedins, I. (2014). *A multi-class multi-server accumulating priority queue with application to health care*. Operations Research for Health Care, 3(2), 73-79.
91. Shi, P.; Chou, M.C.; Dai, J.G.; Ding, D.; Sim, J. (2013). *Hospital inpatient operations: Mathematical models and managerial insights*. Working Paper, Georgia Institute of Technology.
92. Siddharthan, K.; Jones, W.J.; Johnson, J.A. (1996). *A priority queuing model to reduce waiting times in emergency care*. International Journal of Health Care Quality Assurance, 9(5), 10-16.
93. Silberholz, J.; Anderson, D.; Golden, B.; Harrington, M.; Hirshon, J.M. (2013). *The impact of the residency teaching model on the efficiency of the emergency department at an academic center*. Socio-Economic Planning Sciences, 47(3), 183-190.

94. Silva, C.R.N. (2005). *Aplicação de modelos de redes de filas abertas no projeto e planejamento de sistemas discretos de manufatura*. São Carlos, tese de doutorado, PPG-EP/UFSCar, 272pg.
95. Silva, C.R.N.; Morabito, R. (2007). *Aplicação de modelos de redes de filas abertas no planejamento do sistema job-shop de uma planta metal-mecânica*. Gestão & Produção, 14(2), 393-410.
96. Singer, M.; Donoso, P. (2008). *Assessing an ambulance service with queuing theory*. Computers & Operations Research, 35(8), 2549-2560.
97. Sprivulis, P.C.; Da Silva, J.A.; Jacobs, I.G.; Jelinek, G.A.; Frazer, A.R. (2006). *The association between hospital overcrowding and mortality among patients admitted via Western Australian emergency departments*. Medical Journal of Australia, 184(5), 208-212.
98. Stout, W.A.; Tawney, B. (2005). *An Excel forecasting model to aid in decision making that effects hospital bed/resource utilization—hospital capability to admit emergency room patients*. In: Bass EJ, editor. Proceedings of the 2005 systems and information engineering design symposium.
99. Suri, R.; Sanders, J.L.; Kamath, M. (1993). *Performance Evaluation of Production Networks*. Handbooks in Operations Research and Management Science, Vol. 4: Logistics of Production and Inventory, edited by S. C. Graves, A. H. G. Rinnooy Kan, and P. H. Zipkin, Elsevier, Amsterdam, The Netherlands, 199-286.
100. Taylor, I.D.S.; Templeton, J.G.C. (1980). *Waiting Time In a Multi-Server Cutoff-Priority Queue, and Its Application to an Urban Ambulance Service*. Operations Research 28(5):1168-1188.
101. Vaghani, K.; Thakkar, V.; Vaghasiya, S.; Thaker, J.; Bhise, A. (2024). *Implementation of Queuing Theory in Emergency Departments*, 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), Gwalior, India, pp. 1-6, doi: 10.1109/IATMSI60426.2024.10503130.
102. Vass, H.; Szabo, Z.K. (2015). *Application of queuing model to patient flow in emergency department*. Case study. Procedia Economics and Finance, 32, 479-487.
103. Wang, J.; Li, J.; Howard, P.K. (2013). *A system model of work flow in the patient room of hospital emergency department*. Health CareManage. Sci.,16(4), 341–351.
104. Wein, L.M. (1990b). *Scheduling Networks of Queues: Heavy Traffic Analysis of a Two-Station Network with Controllable Inputs*. Operations Research, Vol. 38, No. 6, 1065-1078.
105. Welch, S.J.; Asplin, B.R.; Stone-Griffith, S.; Davidson, S.J.; Augustine, J.; Schuur, J. (2011). *Emergency department operational metrics, measures and definitions: Results of the second Performance measures and benchmarking summit*. Ann. Emerg. Med., 58(1),33–40.
106. Whitt, W. (2005). *Engineering solution of a basic callcenter model*. Manage. Sci., 51, 221–235.
107. Wiler, J.L.; Bolandifar, E.; Griffey, R.T.; Poirier, R.F.; Olsen, T. (2013). *An emergency department patient flow model based on queueing theory principles*. Acad. Emerg. Med., 20(9), 939–946.
108. Wiler, J.L.; Griffey, R.T.; Olsen, T. (2011). *Review of modeling approaches for emergency department patient flow and crowding research*. Acad. Emerg. Med., 18(12):1371-9. doi: 10.1111/j.1553-2712.2011.01135.x. PMID: 22168201.
109. Wu, X.; Xu, R.; Li, J.; Khasawneh, M.T. (2019). *A simulation study of bed allocation to reduce blocking probability in emergency departments: a case study in China*. Journal of the Operational Research Society, 70(8):1376–1390. https://doi-org.ez27.periodicos.capes.gov.br/10.1080/01605682.2018.1506430

110. Yankovic, N.; Green, L.V. (2011). *Identifying good nursing levels: A queuing approach*. Oper. Res.,59(4), 942–955.
111. Yom-Tov, G.B.; Mandelbaum, A. (2014). *Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing*. Manufacturing & Service Operations Management, 16(2), 283-299.
112. Zeltyn, S.; Marmor, Y.N.; Mandelbaum, A.; Carmeli, B.; Greenshpan, O.; Mesika, Y.; Basis, F. (2011). *Simulation-based models of emergency departments: Operational, tactical, and strategic staffing*. ACM Transactions on Modeling and Computer Simulation (TOMACS), 21(4), 1-25.
113. Zhang, A.; Zhu, X.; Lu, Q.; Zhang, R. (2019). *Impact of prioritization on the outpatient queuing system in the emergency department with limited medical resources*. Symmetry 11(6): 796–810. https://doi-org.ez27.periodicos.capes.gov.br/10.3390/sym11060796.
114. Zonderland, M.E.; Boucherie, R.J.; Carter, M.W.; Stanford, D.A. (2015). *Modeling the effect of short stay units on patient admissions*. Operations Research for Health Care. 5, 21-27.