

Metodologia de análise de rede de filas multi-classes para melhorar o desempenho de serviços de urgência

Alécio Rodrigues

Programa Pós-Graduação em Engenharia de Produção – Univ. Federal de Minas Gerais (UFMG)
Av. Presid. Antônio Carlos, 6627 Pampulha - Belo Horizonte, MG, Brasil
alecior@ufmg.br

Lásara F. Rodrigues

Departamento de Engenharia de Produção – Universidade Federal de Minas Gerais (UFMG)
Av. Presid. Antônio Carlos, 6627 Pampulha - Belo Horizonte, MG, Brasil
lasarafrodrigues@gmail.com

Maurício C. de Souza

Departamento de Engenharia de Produção – Universidade Federal de Minas Gerais (UFMG)
Av. Presid. Antônio Carlos, 6627 Pampulha - Belo Horizonte, MG, Brasil
prof.mauriciodesouza@gmail.com

RESUMO

Este trabalho deriva um modelo de rede de filas aberta de um projeto de serviço de urgência destinado a aumentar a capacidade do pronto-socorro para tratar seus pacientes. 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 resultantes é utilizada, para dimensionar as áreas do pronto-socorro usando o tempo de espera, fator de utilização e a probabilidade de *overflow* como metas de qualidade de serviço. Padrões de chegada não homogêneos, distribuições não exponenciais de tempo de atendimento e cinco perfis de pacientes são todos incorporados. O modelo foi aplicado em um pronto-socorro de um hospital privado.

Palavras-chave: Redes de filas abertas, curvas de *trade-off*, otimização.

ABSTRACT

This work derives an open queuing network model from an emergency department project aimed at increasing the emergency department's capacity to treat its patients. The model captures hospital-specific differences in the clinical risk profile of its patients, arrival patterns and volumes, and process efficiency in a single computational model. An implementation of the resulting queuing equations is used, to size the emergency room areas using waiting time, utilization rate and overflow probability as service quality goals. Non-homogeneous arrival patterns, non-exponential service time distributions, and multiple patient profiles are all incorporated. The model was applied in an emergency department of a private hospital.

Keywords: Open Queue network, trade-off curve, optimization.

1. Introdução

A superlotação dos prontos-socorros é 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 erros médicos [Gordon et al. 2001], falta de comunicação e estresse, bem como menor produtividade e moral.

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, segundo as projeções populacionais do IBGE, é de dobrar a taxa de idosos, alcançando 20% da população, feito este que a França levou 150 anos para alcançar. Outra questão a ser observada, 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, também apresentam índices bastante relevantes. 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.

Já pelo lado da oferta, 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, conforme levantamento do Cadastro Nacional de Estabelecimentos de Saúde – CNES. Portanto, o cenário é complexo e a solução deve ser a de melhorar a eficiência dos recursos disponíveis. Desta forma, o objetivo deste trabalho é promover o alinhamento entre a capacidade instalada dos recursos (locais de cuidado e profissionais) do pronto-socorro e a demanda dos pacientes, observando a qualidade do serviço prestado, bem como a segurança do paciente.

Este trabalho está organizado em seções, assim descritas: 2. Revisão bibliográfica sobre a teoria de filas aplicada à otimização do fluxo de pacientes; 3. Descrição do problema da superlotação enfrentado pelos serviços de urgência; 4. Modelos e métodos de avaliação de uma rede de filas; 5. O serviço de urgência como uma rede de filas aberta (OQN); 6. Descrição do método proposto e seus resultados para análise de rede de filas multi-classes; 7. Conclusão e; 8. Referências bibliográficas.

2. Revisão bibliográfica

A teoria de filas tem tido um papel de destaque nas pesquisas relacionadas à otimização do fluxo de pacientes. [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.

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.

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).

[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. [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.

3. Descrição do problema

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, provocando longas esperas (*boarding*). Estudos têm demonstrado que a superlotação não é um fenômeno sem consequências e muitos destes trabalhos têm demonstrado um aumento de erros de procedimentos nos serviços que apresentam este tipo de problema e pior, há uma forte correlação entre o tempo de permanência do paciente “internado” no pronto-socorro e o seu desfecho, seja por aumento do tempo de permanência global como por aumento da mortalidade. 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 de erros médicos (50% das causas de erros num hospital).

4. 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 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.

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 (roteiro) ao longo das estações, determinística ou probabilística, pode ser sequencial, sequencial com realimentação (*feedback*), arborescente, acíclica e cíclica.

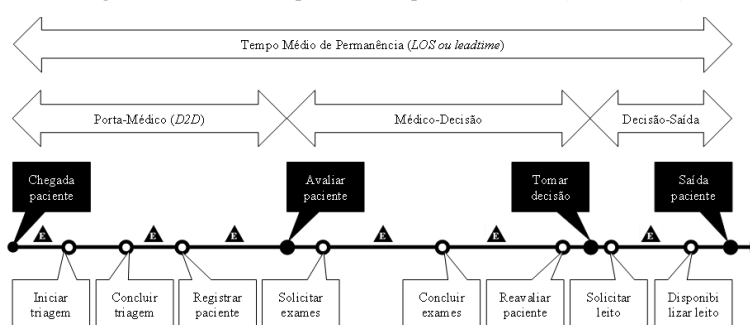
O modelo de rede de filas analisado neste trabalho refere-se a *OQN* (modelos de *OQN*, além de serem analiticamente mais tratáveis do que modelos de *CQN*, podem ser utilizados para aproximá-los (veja [Whitt 1984] e [Calabrese 1992]). Os modelos pressupõem que o sistema atinja o estado de equilíbrio 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 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.

5. O serviço de urgência como uma rede de filas abertas

O pronto-socorro do hospital X, 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 1.

Figura 1 – Jornada do paciente no pronto-socorro (fonte: autor)



Portanto, a jornada do paciente no pronto-socorro 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, internacionalmente reconhecido, 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%. 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*). A classificação, pelo protocolo, também determina o tempo de espera (priorização) para atendimento do paciente, conforme Tabela 1:

Tabela 1 – Protocolo de Manchester

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

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 2 apresenta os principais indicadores utilizados em serviços de urgência:

Tabela 2 – Indicadores de processo mais 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 PS, uma vez que, a taxa de evasão e o índice de satisfação do paciente estão diretamente relacionados com ele. Dentro deste intervalo estão inclusos os tempos de registro e de classificação do risco do paciente, bem como as esperas entre estes processos.
Tempo	Corresponde ao intervalo de tempo, em minutos, entre o início da consulta médica e a decisão médica pela

médico-decisão	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 Lab./Rad. (TurnAroundTime)	Intervalo de tempo 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 (<i>boarding</i>) 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.
Tempo médio de permanência (LOS ou leadtime)	Também conhecido como <i>LOS</i> ou <i>leadtime</i> , corresponde ao intervalo de tempo entre a chegada do paciente até a sua saída física do PS (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 PS.
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 do mesmo período. Este indicador é sensível ao tempo porta-médico.
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 <i>NPS</i> para avaliar serviços hospitalares, dado a sua simplicidade, confiabilidade e flexibilidade. Consiste em uma métrica de lealdade do cliente, criada por [Frederick Reichheld 2003] com o objetivo de medir o grau de lealdade dos clientes das empresas, de qualquer segmento, trazendo reflexos da sua experiência e satisfação. O <i>NPS</i> 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 de 0 a 10. O <i>NPS</i> é calculado subtraindo a porcentagem entre Detratores (0 a 6) e Promotores (9 ou 10). Os Neutros contam para o número total de entrevistados, mas não afetam diretamente o resultado global.

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 3.

Tabela 3 – Indicadores PS Hosp. X e referência EDBA (Dark C. *et al.*, 2020)

Indicadores de processo	PS Hosp. X	EDBA
Tempo médio de permanência (<i>LOS</i>), min.	230	140
Tempo porta-médico (<i>D2D</i>), min.	65	16
Tempo médico-decisão, min.	140	-
Tempo (<i>TAT</i>) de resposta Lab./Rad., min.	120	-
Tempo decisão-saída (alta), min.	15	-
Tempo de <i>boarding</i> , min.	60	71
Taxa de evasão (<i>LWBS</i>), %	2,2	1,5
Taxa de conversão, %	6,25	11,00
Índice de satisfação do cliente (<i>NPS</i>)	37	-

O tempo porta-médico apurado foi de 65 minutos, considerado alto quando comparado com outros prontos-socorros de mesmo porte, conforme referência EDBA na Tabela 3 [Dark C. *et al.* 2020]. E, por esta razão, a taxa de evasão (*LWBS*), é também considerada elevada para os padrões, tendo alcançado 2,2%. Sendo estes, os dois indicadores que mais influenciam na avaliação da experiência do usuário (*NPS*), a quantidade de detratores e de neutros levaram a instituição a uma baixa avaliação (37 pontos), o que a coloca em uma “zona de aperfeiçoamento”.

O segundo momento (médico-decisão) também não responde satisfatoriamente e compromete a jornada do paciente. O tempo ponderado entre os pacientes em alta e internação, alcançou 140 minutos. Neste intervalo de tempo estão inclusos a consulta médica, a realização de exames de laboratório e/ou imagem e a reavaliação médica, conforme Tabela 3. 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. Ou seja, elevada demanda por exames e tempos de resposta também relativamente altos, comprometem consideravelmente 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 3). Tempo ótimo

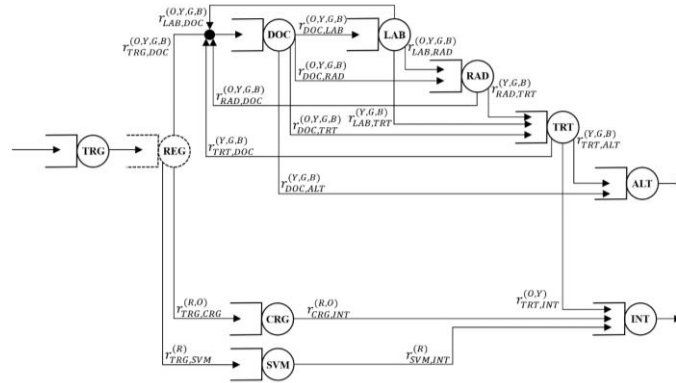
quando comparado com a referência EDBA na própria Tabela 3, mas que neste caso poderia ser ainda melhor em função da subutilização ($\rho=47\%$) dos leitos nas unidades de internação.

Desta forma, o tempo médio ponderado de permanência do paciente no pronto-socorro do hospital X foi de 230 minutos ou 3 horas e 50 minutos. Tempo sido considerado alto, quando comparado (ver Tabela 3) com os hospitais referência EDBA, de igual porte. Entretanto, devido ao elevado percentual de pacientes considerados de baixo risco clínico ($>85\%$) este tempo pode ser melhorado consideravelmente.

Observando a rede de filas de pacientes no pronto-socorro, ilustrado na Figura 2, 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 ser 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: apenas uma consulta médica e saírem em alta do pronto-socorro ou, além da consulta médica, podem realizar exames de laboratório (*LAB*) e/ou exames de imagem (*RAD*) e/ou tratamento (*TRT*) (medicação, observação, entre outros procedimentos). A grande maioria destes, deixam o PS em alta (*ALT*) e, apenas uma pequena parcela, somada àqueles provenientes do *CRG* e *SVM*, seguirão para unidades de internação (*INT*).

Figura 2 – Priorização (*as-is*) dos pacientes nos 9 nós do pronto-socorro onde $r_{i,j}^{(t)}$ representa a probabilidade de fluxo do nó *i* para *j* conforme o risco *t* do paciente

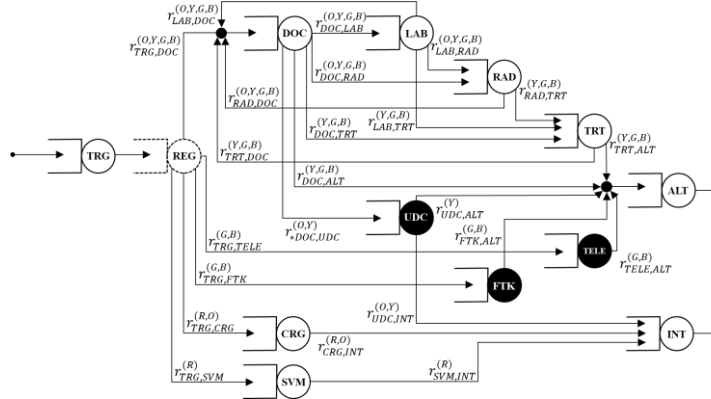


6. Análise de rede de filas multi-classes

A Figura 3 apresenta a rede de filas proposta neste trabalho para analisar a capacidade de cada uma das áreas do pronto-socorro, com a inclusão de três novas, conforme ilustrado e destacado na Figura 3: o *fast-track* (*FTK*), um consultório de fluxo rápido dedicado aos pacientes de baixo risco clínico, a tele consulta (*TELE*), um consultório de fluxo igualmente rápido para pacientes, também, de baixo risco clínico, podendo ser acessado de forma presencial no PS ou remotamente e, a unidade de decisão clínica (*UDC*), uma área dedicada a pacientes cuja decisão médica demanda maior tempo, porém, limitada em até 4 horas.

A segmentação, conforme Figura 3, consiste em desviar parte do fluxo de pacientes ($r_{TRG,FTK}^{(G,B)}$ e $r_{TRG,TELE}^{(G,B)}$) que inicialmente iriam para o consultório médico (*DOC*), representado na Figura 3 pelo fluxo $r_{TRG,DOC}^{(O,Y,G,B)}$. Outra modificação proposta, foi a adoção de uma unidade de decisão clínica (*UDC*) para limitar o tempo da decisão médica para aqueles pacientes, sobretudo os urgentes (amarelos), cuja decisão é relativamente mais difícil.

Figura 3 – Segmentação (*to-be*) dos pacientes nos 12 nós do pronto-socorro onde $r_{i,j}^{(t)}$ representa a probabilidade de fluxo do nó i para j conforme o risco t do paciente (fonte: autor)

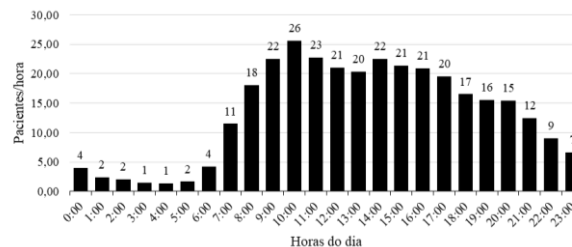


Para cada uma das etapas do modelo, há uma seção correspondente que deriva seus métodos e explica sua notação matemática. Ao longo deste trabalho, foram utilizadas notações e variáveis baseadas em filas comumente aceitas.

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 4), os tempos médios de 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 4 – Taxa média de chegada por hora no pronto-socorro (fonte: hospital X)



Ao chegar, os pacientes são classificados através do protocolo de Manchester, já utilizado pelo hospital e seus volumes podem ser mensurados a partir dos dados históricos do hospital. Para calibrar o nível geral de demanda na rede de fluxo de pacientes, foi utilizado um volume anual histórico e, a partir deste, foram considerados os efeitos de sazonalidade e de pico.

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, quando nos meses em que o volume ficou próximo de 12.000 pacientes a taxa de evasão oscilou entre 3% e 4%. Lembrando que, a taxa de evasão (*LWBS*), bem como o índice de satisfação do paciente (*NPS*) são sensíveis ao tempo porta-médico (*D2D*).

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.

Os pacientes saem do serviço de urgência com base em frações de pacientes admitidos (*INT*) no hospital ou frações de pacientes em alta. 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.

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 tele consulta, 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 tele consulta pode, inclusive, ser acessado de forma remota, evitando o deslocamento do paciente até o pronto-socorro.

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

$$r_{ij}^{(R)} = \begin{bmatrix} \dots & 0 & 0 & 0 & 0,0385 & 0,9616 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0,0385 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0,9616 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

$$r_{ij}^{(O)} = \begin{bmatrix} \dots & 0,9696 & 0 & 0 & 0,0304 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \dots & 0 & 0 & 0 & 0 & 0,2930 & 0,2000 & 0 & 0,9696 & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0,0304 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0,2930 & 0 & 0 & 0 & 0 & \dots & 0,0310 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0,2310 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0,9696 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

A matriz-r para os pacientes cujo perfil é urgente (Y), consiste naquela de maior complexidade, com maiores conexões. As probabilidades de segmentação dos perfis de risco, pouco urgente (G) e não urgente (B) também obedecem aos mesmos princípios dos perfis anteriores.

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]. A equação (1) é uma fórmula geral para converter um volume anual da emergência em chegadas de pacientes por hora usando picos horários (índice: 1,50) e/ou sazonalidades (índice: 1,18):

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

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). As taxas de chegada aos outros nós são limitadas às transferências internas para este nó. Portanto, para todos os nós, exceto o nó da triagem, a equação (2) torna-se equação (3) em nossa rede de 12 nós com cinco tipos de pacientes:

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

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 Tabela 4.

Tabela 4 – Fluxo de paciente em cada um dos 12 nós (i)

λ_i	Pts./h
$\lambda_{TRG} = 23,4927 \times (1,0000) =$	23,49
$\lambda_{DOC} = 23,4927 \times (0,0110 \times 1,4936 + 0,1278 \times 1,4840 + 0,6483 \times 0,4102 + 0,2521 \times 0,4096) =$	13,14
$\lambda_{TELE} = 23,4927 \times (0,6483 \times 0,1500 + 0,2125 \times 0,1500) =$	3,03
$\lambda_{FTK} = 23,4927 \times (0,6483 \times 0,5500 + 0,2125 \times 0,5500) =$	11,12
$\lambda_{CRG} = 23,4927 \times (0,0003 \times 0,0385 + 0,0110 \times 0,0304) =$	0,01
$\lambda_{SVM} = 23,4927 \times (0,0003 \times 0,9616) =$	0,01
$\lambda_{LAB} = 23,4927 \times (0,0110 \times 0,2930 + 0,1278 \times 0,2930 + 0,6483 \times 0,0879 + 0,2125 \times 0,0879) =$	2,73
$\lambda_{RAD} = 23,4927 \times (0,0110 \times 0,2310 + 0,1278 \times 0,2310 + 0,6483 \times 0,0693 + 0,2125 \times 0,0693) =$	2,16
$\lambda_{TRT} = 23,4927 \times (0,1278 \times 0,1570 + 0,6483 \times 0,1500 + 0,2125 \times 0,1500) =$	3,51
$\lambda_{UDC} = 23,4927 \times (0,0110 \times 0,9696 + 0,1278 \times 0,2000) =$	0,85
$\lambda_{ALT} = 23,4927 \times (0,1278 \times 0,5000 + 0,6483 \times 1,0000 + 0,2125 \times 1,0000) =$	21,72
$\lambda_{INT} = 23,4927 \times (0,0003 \times 1,0000 + 0,0110 \times 1,0000 + 0,1278 \times 0,4000) =$	1,47

e) Estabelecer a qualidade do serviço

Especificar a qualidade do serviço em termos de tempos de espera do paciente (W_q^{meta}), taxa de utilização (ρ^{meta}) e probabilidades de estouro ou *overflow* da área (p_c^{meta}) é mais intuitivo para os gestores deste tipo de sistema. 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 *overflow* explicam melhor a experiência do paciente e sensibilizam os gestores de pronto-socorro. Ver Tabela 5.

Tabela 5 – Qualidade requerida do serviço

Área_i	P_{c_i}	w_{q_i}	ρ^{meta}
Triagem, TRG		≤ 5 min.	
Consulta Médica, DOC		≤ 15 min.	$0,70 < \rho \leq 0,80$
Tele Consulta, $TELE$		≤ 5 min.	
Fast-Track, FTK		≤ 5 min.	$0,70 < \rho \leq 0,80$
Centro Cirúrgico, CRG	$< 2\%$		
Sala Vermelha, SVM	$< 5\%$	0 min.	
Laboratório, LAB		≤ 15 min.	
Radiologia, RAD		≤ 15 min.	
Tratamento, TRT			
Unid. Decisão Clínica, UDC		≤ 15 min.	$0,70 < \rho \leq 0,80$
Alta Médica, ALT		≤ 15 min.	$0,70 < \rho \leq 0,80$
Unidades Internação, INT	$< 5\%$	≤ 60 min.	$0,70 < \rho \leq 0,80$

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. O número de servidores (c), profissionais ou locais de cuidado, dependendo da área em análise, é calculado utilizando a Equação (4), onde λ é a taxa de chegada (em pacientes/hora) e μ a taxa de serviço (em pacientes/hora):

$$c = \frac{\lambda}{\mu \times \rho^{meta}} \quad (4)$$

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 :

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

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 de ciclo (médio). A p_0 para uma fila $M/M/c$ foi calculada, através da equação (6), como:

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

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).

$$P_c = \frac{\left(\frac{\lambda}{\mu} \right)^c}{c!} \frac{1}{\sum_{i=0}^c \left(\frac{\lambda}{\mu} \right)^i \frac{1}{i!}} \quad (7)$$

Tabela 6 – Capacitação dos nós

Área _i	$c^{instalado}$	$c^{projetado}$	P_{c_i}	w_{q_i}	ρ
Triagem, TRG	2	2	32,3%	03,97'	0,78
Registro, REG (nó virtual)	3	3	26,1%	01,94'	0,78
Consulta Médica, DOC	12	8	08,2%	02,98'	0,66
Tele Consulta, TELE	0	1	37,8%	04,63'	0,61
Fast-Track, FTK	0	4	22,6%	08,27'	0,79
Centro Cirúrgico, CRG	1	1	01,6%	02,00'	0,02
Sala Vermelha, SVM	2	2	00,0%	00,01'	0,01
Laboratório, LAB	-	4	12,5%	06,56'	0,57
Radiologia, RAD	2	2	08,6%	00,61'	0,27
Tratamento, TRT	30	13	0,00%	00,01'	0,27
Unid. Decisão Clínica, UDC	0	8	01,5%	00,35'	0,43
Alta Médica, ALT	0	6	14,3%	02,70'	0,72
Unidades Internação, INT	207	207	00,0%	12,65'	0,46

Dado o volume de pacientes e as premissas de qualidade requerida, para cada uma das áreas (nós) do serviço de urgência, conforme Tabela 6, através das equações (4), (5), (6) e (7) foram calculados: a necessidade de profissionais ou locais de cuidado ($c^{projetado}$), conforme o caso, a probabilidade de overflow (P_c), o tempo de espera (W_q) e a utilização (ρ).

g) Calcular o tempo porta-médico

A etapa final do modelo 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. Na segmentação do fluxo de pacientes do pronto-socorro, há 2 processos comuns aos 5 perfis de pacientes, o de classificação de risco (TRG) o de registro (REG), cujos tempos de espera (na fila) e de ciclo (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):

$$T_{D2D} = (W_q(\text{TRG}) + T_{\text{ciclo}}(\text{TRG})) + (W_q(\text{REG}) + T_{\text{ciclo}}(\text{REG})) + \left[\frac{\lambda_{\text{DOC}}}{\lambda_{\text{TRG}}} \times (T_{\text{desloc}} + W_q(\text{DOC})) \right] \quad (8)$$

$$+ \left[\frac{\lambda_{\text{TELE}}}{\lambda_{\text{TRG}}} \times (T_{\text{desloc}} + W_q(\text{TELE})) \right] + \left[\frac{\lambda_{\text{FTK}}}{\lambda_{\text{TRG}}} \times (T_{\text{desloc}} + W_q(\text{FTK})) \right]$$

$$T_{D2D} = (3,97 + 4,00) + (1,94 + 6,00) + \left[\frac{13,14}{23,49} \times (5,00 + 2,98) \right] + \left[\frac{3,03}{23,49} \times (5,00 + 4,63) \right] + \left[\frac{11,12}{23,49} \times (5,00 + 8,27) \right]$$

$$\cong 28 \text{ min.}$$

Todos os termos na equação (8) foram calculados anteriormente, conforme Tabelas 4 (λ) e 6 (W_q), exceto os tempos de deslocamento (T_{desloc}), que foram estimados em 5 minutos, independente da classificação de risco e do fluxo do paciente.

7. Conclusão

A metodologia aqui proposta, desenvolvida em Python, 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 PS. 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.

A transformação não apenas aprimorou os indicadores de desempenho, conforme destacado na Tabela 7, mas também redefiniu o padrão de excelência, proporcionando um ambiente mais seguro e eficiente para cuidados emergenciais de saúde, além de torná-lo uma referência para outros serviços de urgência.

Tabela 7 – Ganho potencial com implantação da metodologia

Indicadores	As-is	To-be	Resultado
Tempo médio de permanência (LOS), min.	230	≤ 170	-26%
Tempo porta-médico (D2D), min.	65	≤ 30	-54%
Tempo médico-decisão, min.	140	120	-14%
Tempo de boarding, min.	60	≤ 15	-75%
Taxa de evasão (LWBS), %	2,2	< 1,4	-36%
Índice de satisfação do cliente (NPS)	37	≥ 75	+103%
Volume, mil pts./ano	112	156	+39%

8. Referências bibliográficas

1. Bitran, G.R.; Sarkar, D. (1994). *Targeting Problems in Manufacturing Queueing Networks - An Iterative Scheme and Convergence*. European Journal Oper. Res., Vol.76, No.3, 501-510.
2. Calabrese, J.M. (1992). *Optimal Workload Allocation in Open Networks of Multiserver Queues*. Management Science, 38, 12, 1792–1802.
3. Cochran, J.K.; Roche, K.T. (2009). *A multi-class queueing network analysis methodology for improving hospital emergency department performance*. Comp. Oper. Res., 36(5), 1497–1512.
4. 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.
5. 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.
6. Derlet, R.W.; Richards, J.R. (2002). *Emergency department overcrowding in florida, new york, and texas*. Southern medical journal, 95(8), 846-850.
7. Disney, R.L.; Konig, D. (1985). *Queueing Networks: A Survey of Their Random Processes*. SIAM Review, Vol. 27, 335-403.
8. Filippatos, G.; Evridiki, K. (2015). *The effect of Emergency department crowding on patient outcomes*. Health Science Journal 2015; 9(16): 1–6.

9. Fomundam, S.; Herrmann, J. (2007). *A survey of Queuing Theory Applications in healthcare*. ISR Technical Report 24.
10. Gershwin, S.B. (1994). *Manufacturing Systems Engineering*. Prentice Hall, Englewood Cliffs, NJ.
11. Gordon, R., Graber, M., and Franklin, N. (2002). *Reducing diagnostic errors in medicine: what's the goal?*. Academic Medicine, 77(10), 981-992.
12. Green, L.V.; Soares, J.; Giglio, J.F.; Green, R.A. (2006). *Using queuing theory to increase the effectiveness of emergency department provider staffing*. Academic Emergency Medicine;13(1):61-68.
13. Gross, D.; Harris, C.M. (1998). *Fundamentals of queueing theory*. 3rd ed. Section 4.2: Open Jackson networks. New York: Wiley; p. 174-83.
14. Hsu, L.F.; Tapiero, C.S.; Lin, C. (1993). *Network of Queues Modeling in Flexible Manufacturing Systems: A Survey*. Recherche Operat. l'Operations Research, 27(2), 201-248.
15. Huang, J. (2013). *Patient flow management in emergency departments*. Ph.D. Dissertation, National University of Singapore.
16. 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.
17. Kao, E.P.C.; Tung, G.G. (1981). *Bed allocation in a public healthcare delivery system*. Manage. Sci., 27(5), 507-520.
18. 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.
19. 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.
20. Lakshmi, C.; Sivakumar, A.I. (2013). *Application of queueing theory in healthcare: a literature review*. Oper. Res. Healthc. 2. 25-39.
21. 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.
22. Ozcan, Y.A. (2005). *Forecasting*. In: *Quantitative methods in health care management*, California: Josey-Bass; 2005. p. 10-44 [Chapter 2].
23. Roche, K.T.; Cochran, J.K. (2007). *Improving patient safety by maximizing fast-track benefits in the emergency department: A queueing network approach*. Proceedings of the 2007 Industrial Engineering Research Conference, 619-624.
24. 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.
25. Silva, C.R.N.; Morabito, R. (2007). *Aplicação de modelos de redes de filas abertas no planejamento do sistema jobshop de uma planta metalmecânica*. Gest&Prod, 14(2), 393-410.
26. 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.
27. 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.
28. 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.
29. 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.