Sistema de Filas Infinitas

Aplicações

Micael Egídio Papa da Silva - 211029236
Universidade de Brasília

Julho, 2023

1

Introdução

Um processo de fila é um modelo matemático usado para descrever o comportamento de sistemas de espera, em que o cliente chega e entra na fila para serem atendidos pelos servidores disponíveis. Esse modelo é utilizado em várias áreas como logística, atendimento ao cliente e transporte. Esses tipos de processos precisam de elementos como:

1. Clientes: Representam entidades que querem algum tipo de serviço ou atendimento.
2. Fila: Local onde os clientes aguardam o atendimento ou serviço enquanto os servidores estão ocupados.
3. Servidores: Aqueles responsáveis pelo atendimento ou serviço.
4. Política de Atendimento: Como os servidores selecionam os clientes quando estão livres, podendo ser por ordem de chegada, ordem de prioridade etc.
5. Distribuição de chegada e distribuição de serviço: São distribuições usadas para modelar/descrever os padrões de chegada dos clientes e do atendimento prestado por cada servidor.

Ao utilizar um processo de fila, é possível analisar diversas métricas, como o tempo médio de espera na fila, o tempo médio de atendimento, a taxa de ocupação dos servidores e a probabilidade de encontrar a fila cheia.

Os processos de fila são usados para otimizar a eficiência dos sistemas, realocar recursos da melhor forma e melhorar a qualidade de serviço prestado aos clientes. Tais processos fornecem uma visão quantitativa do sistema, permitindo tomar decisões para melhorar o fluxo de atendimento e reduzir o tempo de espera dos clientes.

Os processos de filas com infinitos servidores são um tipo especial de sistema de filas em que o número de servidores disponíveis para atender os clientes é considerado infinito. A suposição de serem infinitos os servidores permite trazer uma análises mais precisas em relação as métricas e, principalmente, quanto ao tempo médio de espera e a taxa de ocupação.

2

Processos de Fila com Infinitos Servidores

A esfera científica optou por padronizar a descrição dos sistemas de filas em termos da notação de David George Kendall (KENDALL, 1953) , que consiste em uma série de símbolos separados por barras, A/S/C/R/Q, onde A caracteriza a distribuição do tempo de chegada, S a distribuição dos servidores, C é o número de canais disponíveis para atender aos clientes da fila, R a restrição da capacidade do sistema e Q é o comportamento da fila. Os dois últimos são omitidos quando $R = \infty$ (ausência de restrição) e Q= FIFO (first-in-first-out). A notação a ser adotada para infinitos servidores será M/M/∞. , ou seja, ambos os tempos de chegada e o tempo de serviço serão modelados por uma distribuição exponencial proveniente das propriedades markovianas e o número de servidores será ∞.

Na fila M/M/∞, os usuários chegam aos servidores de acordo com um processo de Poisson com média λ (implicando no tempo de chegada ser exponencialmente descrito com média $\frac{1}{\lambda}$), e os tempos dos servidores são independentes das chegadas e seguem uma distribuição exponencial com média $\frac{1}{\lambda}$. O parâmetro λ denota a taxa de chegada e μ a taxa de serviço.

Apesar do objetivo inferencial algumas vezes optar por focar na estimação de parâmetros que caracterizam o sistema (i.e λ e μ), é mais comum optar por observar termos quantitativos que descrevam a utilização, eficiência e congestionamento dentro de um cenário de sistema de filas. Diferentemente das filas M/M/c, uma M/M/∞ sempre alcançará sua estacionariedade para todo λ e μ. Na maioria dos casos o interesse será puramente prever, dentro de um estado estacionário, o número de usuários do sistema, N, e do tempo que o usuário permanece dentro do sistema, W (tempo de espera).

Em uma fila M/M/∞ em equilíbrio, a distribuição do (estado-estacionário) numero de usuários no sistema, N, será uma Poisson com parâmetro $\frac{\lambda}{\mu}$:

$$p(N = n|\mu, \lambda) = \frac{\left(\frac{\lambda}{\mu}\right)^n e^{\lambda/\mu}}{n!}, n = 0, 1, 2, ... \tag{2.1}$$

Para essa fila, N também é o número de servidores ocupados (outra possível medida de performance). O termo $\frac{\lambda}{\mu}$ (média de chegadas por unidade de tempo quando a unidade temporal assume a média de serviços) é denotada como "carga oferecida", sendo uma quantidade adimensional

cuja unidade é o "erlang". Assumiremos $\frac{\lambda}{\mu}$ como θ. Perceba que em 2.1, ou seja , em uma fila M/M/∞ a carga oferecida também será a expectância de N , sendo assim interpretada como o número médio de servidores que a população de usuários "almeja"administrar simultaneamente (MITRANI, 1992).

Como a fila M/M/∞ presume que p usuário seja atendido sem nunca promulgar a fila, o tempo em que o usuário passa no sistema W será evidentemente igual ao tempo demandado pelo próprio usuário para ser atendido, logo, a distribuição do tempo de espera será dada através de um servidor exponencialmente distribuído, ou seja:

$$p(w|\mu,\lambda) = p(w|\mu) = \mu e^{-\mu w}, w > 0 \tag{2.2}$$

Vale ressaltar que 2.1 também é válida para uma fila do tipo M/G/∞, onde G indica uma distribuição geral para o tempo de chegada.

Sob a ótica de um comportamento transiente, diferentemente do que ocorre na maioria dos sistemas de filas, probabilidades transientes podem ser derivadas de uma forma fechada para uma fila M/M/∞. De fato, a distribuição do número de usuários no sistema em um tempo t, $N(t)$, é uma Poisson com média $\frac{\lambda(1-e^{-\mu t})}{\mu}$, onde :

$$p(N(t) = n|\mu,\lambda) = \frac{1}{n!}\left[\frac{\lambda(1-e^{-\mu t})}{\mu}\right]^n exp\left(\frac{-\lambda(1-e^{-\mu t})}{\mu}\right) \tag{2.3}$$

Note que 2.3 é válida para sistemas M/G/∞, até mesmo para tempos de chegadas dependentes e e distribuição de tempo de servidores dependentes do tempo (NEWELL, 2013).Pode -se perceber que as probabilidades de 2.3 são derivadas de acordo com o pressuposto usual de que o sistema está vazio no tempo t=0 , ou seja, podemos obter 2.1 de 2.3 ao deixar t ir para ∞.

Dentre outras características a serem ressaltadas acerca da filas M/M/∞ podemos citar:

- As chegadas são aleatórias (como esperado em muitas circunstancias naturais)
- A taxa de chegada é constante (uma simplificação que tende a subestimar a variabilidade produzida pelo sistema)
- Tempo de servidores de diferentes usuários são independentes, de qualquer distribuição.

Modelos de filas infinitas estão sendo desenvolvidos em muitas direções desde suas ideias embrionárias. Refletindo nos 50 anos de estudo acerca das modelagens de filas, Worthington descreve com mais detalhes como os pressupostos das filas M/M/∞ simplificam de forma significante o cunho matemático necessário para a análise de sistemas não-exponenciais , mas um nó unitário e para sistemas com nós múltiplos e para estados de comportamentos estacionários e comportamento de dependência temporal(WORTHINGTON, 2009).

3

Aplicações

Ward Whitt tece comentários acerca da importância das filas M/M/∞ no que se diz respeito ao desenvolvimento da teoria de filas e suas aplicações, apesar do fato de se pressupor infinitos recursos e portanto uma ausência de filas proveniente do imediatismo do atendimento(WHITT, 2016). Whitt também ressalta que as filas de infinitos servidores da uma maior importância para a base das análises da "carga oferecida"(offered load) para sistemas de múltiplos servidores com variação de tempo de chegada. Tal premissa está presente dentro do âmbito da assistência médica (healthcare) onde a habilidade das filas M/M/∞ de modelar a "carga oferecida"ou "demanda irrestrita"se mostrou eficiente. Nesse contexto, tais termos se traduzem essencialmente como o número de pacientes perceptíveis em um sistema em um dado tempo (ou em um nó particular do sistema) se nesse tempo específico o progresso dos pacientes não se atrasa de modo a gerar uma fila de fato. A teoria das aplicações na área da healthcare foram bi-particionadas:

- Mais relacionada as filas M/G/∞ assumindo tempos de chegada modelados por uma Poisson em tempo continuo, cujas ideias foram desenvolvidas por Ward Whitt em conjunto com Eick e Massey (responsáveis por aplicações em networks).

- Possuinte de menos restrições acerca dos tempos de chegada, com suas ideias principais difundidas por Utley,Gallivan, Treasure e Valencia (UTLEY et al., 2003) para sistemas de nó único e com extensões para sistemas de nós múltiplos desenvolvidos por Utley, Gallivan,Pagel e Richards (UTLEY et al., 2009).

Com a difusão dos campos teóricos e o conseguente avanço do conhecimento acerca do comportamento das filas M/M/∞ as aplicações para a assistência médica foi dominada por estudos de incessantes demandas de leitos para os pacientes, seja por uma única ala, para múltiplas alas médicas ou até mesmo para o hospital inteiro, Todavia, existem exemplos que optam por enfatizar na exigência de uma maior capacitação de profissionais , como por exemplo , em departamentos de emergência, socorristas e cuidados comunitários. Modelagens de filas M/M/∞ surgiram da necessidade de oferecer suporte estratégico, tático e operacional para facilitar eventuais tomadas de decisão.

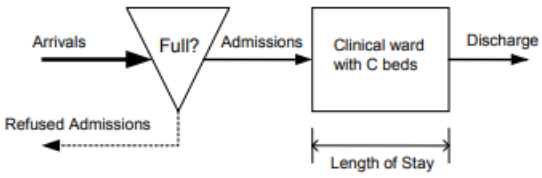


Figura 3.1: Modelo estrutural do fluxo de pacientes em uma enfermaria clínica

Modelando o fluxo de pacientes

Um exemplo de aplicação se deu com R. Bekker e A.M. de Bruin que optaram por investigar em primeiro lugar a eficiência dos tempos de chegada homogêneos de uma Poisson com parâmetro λ restrita ao modelo M/G/c/c que se encarregou por analisar um banco de dados de 24 clinicas durante um período de 3 anos(BEKKER; BRUIN, 2010).

Terminologias e medidas usadas nas aplicações de Healthcare

1. LOS (Length of Stay) : O tempo passado na enfermaria é denominado tempo de permanência, muitas vezes abreviado como LOS, após o qual o paciente recebe alta ou é transferido para outra enfermaria. O LOS é facilmente derivado do sistema de informação do hospital como tempo de admissão e tempo de alta são registrados no nível do paciente individual. Sendo também caracterizado pelas curvas de lorenz e pelos coeficientes de Gini a serem ainda postulados.

2. ALOS (Average length of stay) : é uma estatística autoexplicativa, retorna a média do tempo de permanência do paciente sendo uma unidade de tempo.

3. DSS (Decision support system) : É o suporte estatístico que implementa os resultados obtidos com o auxilio dos processos de filas para a efetivação do realoque de recursos objetivando uma melhora na eficácia do estabelecimento.

4. Curva de Lorenz : A curva de Lorenz é uma representação gráfica da distribuição cumulativa de uma distribuição de probabilidade, onde a porcentagem de pacientes é disposta nas abscissas e a porcentagem de consumo de recursos (em termos de dias de internação) nas ordenadas.

5. Índice de Gini : (denotado como G) é uma medida da dispersão da curva de Lorenz sendo definido como uma razão com valores entre 0 e 1; o numerador é a área entre a curva de lorenz da distribuição e da reta suporte da distribuição uniforme; o denominador é a área da reta suporte distribuição uniforme. Assim, um coeficiente de Gini baixo indica que a variabilidade do LOS é baixa, enquanto um alto coeficiente de Gini indica uma distribuição mais variável. A fórmula a seguir foi usada para calcular o coeficiente de Gini para cada enfermaria clínica:

$$G = \frac{1}{2}\left(n + 1 - 2\frac{\sum_{i=1}^{n}(n+1-i)\,y_i}{\sum_{i=1}^{n} y_i}\right) \tag{3.1}$$

onde : n: número de pacientes admitidos para uma enfermaria y_i : Os valores observados do LOS em ordem crescente , onde $y_i \leq y_{i+1}, i = 1, ..., n-1$

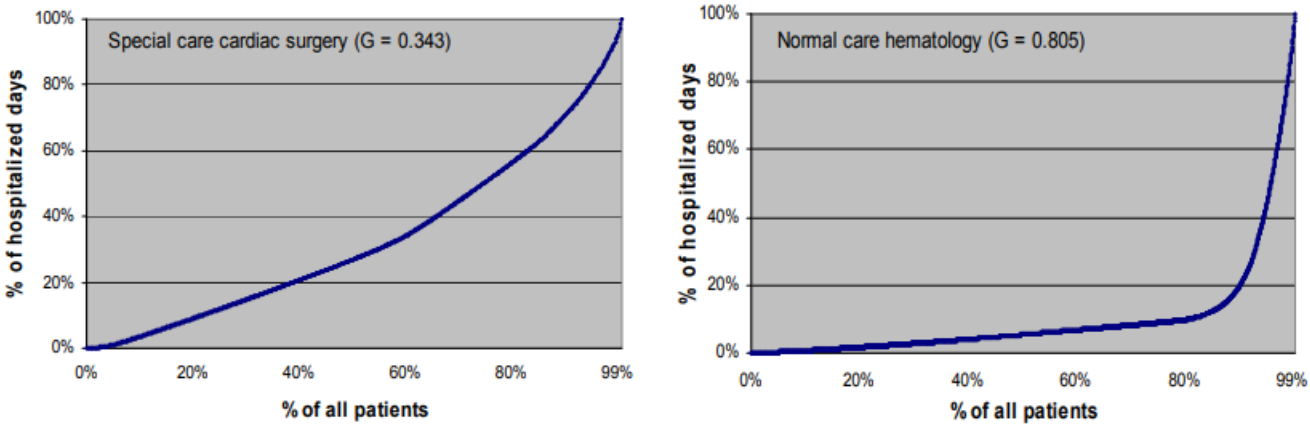


Figura 3.2: Caption

Temos:

Ward description	2004			2005			2006		
	ALOS [days]	C_v [σ/μ]	Gini [G]	ALOS [days]	C_v [σ/μ]	Gini [G]	ALOS [days]	C_v [σ/μ]	Gini [G]
Coronary Care Unit	1.573	1.692	0.638	1.748	1.311	0.564	1.694	1.313	0.569
Intensive Care Unit surgical	5.676	2.690	0.750	6.021	2.012	0.719	5.397	1.730	0.684
Intensive Care Unit medical	4.957	1.851	0.687	5.417	2.025	0.705	5.147	2.018	0.684
Pediatric Intensive Care Unit	3.533	1.997	0.644	4.168	1.571	0.636	4.180	1.512	0.606
Neonatal Intensive Care Unit	9.025	1.721	0.668	8.335	1.985	0.697	7.778	1.644	0.680
Medium Care	2.211	1.478	0.543	2.619	1.738	0.606	2.374	1.643	0.584
Special Care cardiac surgery	1.524	0.733	0.319	1.668	0.644	0.309	1.715	0.821	0.343
NC Cardiac surgery and cardiology	4.005	1.519	0.567	4.054	1.502	0.570	4.347	1.430	0.570
NC Gynaecology	3.444	1.290	0.552	3.615	1.420	0.570	3.172	1.391	0.558
NC Hematology	3.732	2.667	0.819	3.023	2.591	0.812	2.763	2.651	0.805
NC Surgical oncology	6.571	1.110	0.500	7.374	1.111	0.497	6.448	1.167	0.510
NC Internal medicine unit 1	6.493	1.273	0.579	7.266	1.292	0.553	6.354	1.089	0.528
NC Internal medicine unit 2	4.853	1.469	0.653	4.810	1.713	0.683	4.852	1.627	0.685
NC Pediatric unit 1	3.850	1.449	0.588	3.758	1.470	0.591	3.440	1.486	0.579
NC Pediatric unit 2	3.432	1.434	0.599	4.119	2.243	0.633	4.131	1.750	0.630
NC Otolaryngology (ENT)	5.333	1.552	0.612	4.052	1.544	0.577	4.362	1.649	0.596
NC Internal lung	4.765	1.139	0.520	4.517	1.245	0.541	4.633	1.198	0.541
NC Neuro- and orthopedic surgery	6.340	1.527	0.579	5.441	1.472	0.545	5.256	1.302	0.537
NC Neurology	5.921	1.503	0.609	5.597	1.300	0.587	5.533	1.401	0.590
NC Obstetrics	1.589	1.784	0.688	1.438	2.005	0.707	1.501	2.039	0.698
NC Internal oncology	4.904	1.373	0.572	4.061	1.346	0.574	4.527	1.314	0.562
NC Ophthalmology	2.783	1.150	0.499	2.180	1.225	0.487	1.583	1.186	0.534
NC Trauma surgery	7.695	1.299	0.541	7.641	1.251	0.537	6.833	1.175	0.526
NC Vascular surgery	6.195	1.345	0.554	6.844	1.321	0.550	6.487	1.638	0.583
Average	**4.199**	**1.544**	**0.595**	**4.060**	**1.556**	**0.594**	**3.918**	**1.507**	**0.591**

Tabela 3.1: Estatísticas das LOS (2004-2006)

Onde as curvas de Lorenz a serem destacadas serão :

Percebe-se em 3.1 que para 2006 o menor coeficiente de gini foi de 0.343 para a "special care cardiac surgery"e o maior foi 0.805 na "NC hematology"e a 3.2 explicita exatamente os diferentes níveis de curva desses cenários. Essas curvas de Lorenz são **extremamente**

Occupancy rates			
Ward description	**2004**	**2005**	**2006**
Coronary Care Unit	80.7%	79.3%	73.3%
Intensive Care Unit surgical	82.6%	84.1%	76.3%
Intensive Care Unit medical	76.8%	77.5%	71.9%
Pediatric Intensive Care Unit	43.8%	50.2%	44.3%
Neonatal Intensive Care Unit	52.0%	62.1%	67.1%
Medium Care	76.6%	86.8%	69.8%
Special Care cardiac surgery	47.6%	46.3%	48.2%
NC Cardiac surgery and cardiology	78.5%	78.7%	80.8%
NC Gynaecology	63.2%	61.4%	58.6%
NC Hematology	87.4%	83.8%	84.3%
NC Surgical oncology	77.1%	72.1%	79.2%
NC Internal medicine unit 1	88.0%	92.5%	81.1%
NC Internal medicine unit 2	87.0%	91.8%	84.5%
NC Pediatric unit 1	61.3%	64.4%	59.3%
NC Pediatric unit 2	69.3%	80.1%	65.5%
NC Otolaryngology (ENT)	59.1%	67.0%	65.4%
NC Internal lung	73.5%	66.0%	65.8%
NC Neuro- and orthopedic surgery	82.1%	76.3%	72.9%
NC Neurology	63.0%	69.3%	73.0%
NC Obstetrics	41.4%	46.0%	53.9%
NC Internal oncology	74.3%	65.3%	64.1%
NC Ophthalmology	52.0%	58.6%	57.4%
NC Trauma surgery	81.5%	78.9%	78.6%
NC Vascular surgery	75.2%	95.8%	81.0%

Figura 3.3: Taxa de ocupação por enfermaria clínica (2004-2006)

úteis para identificar pacientes que possuam uma estadia prolongada e um consumo desproporcional de recursos Temos que a taxa de ocupação será dada por:

$$\text{Ocupação} = \frac{\text{Número médio de leitos ocupados}}{\text{Número de leitos operacionais}} = \frac{\text{Admissões (por unidade de tempo) x ALOS}}{\text{Número de leitos operacionais}} \tag{3.2}$$

A fração de pacientes que está bloqueada (sem ter acesso ao atendimento) será calculada via :

$$P_c = \frac{\frac{(\lambda\mu)^c}{c!}}{\sum_{k=0}^{c} \frac{(\lambda\mu)^k}{k!}} \tag{3.3}$$

Perceba que esse modelo em particular é insensível À "LOSS-distribution"e é válido para todos os tipos de tempo de servidores , a taxa de ocupação será definida por:

$$\text{Taxa de ocupação} = \frac{(1 - P_c)\,\lambda \cdot \mu}{c} \tag{3.4}$$

Onde o termo $\lambda \cdot \mu$ é denominado como "ofered load"(carga fornecida)

Teremos assim:

Aproximando o número de chegadas

No intuito de aplicar o "Erlang loss model"e para propósitos de validação, existe a necessidade de se quantificar o número de chegadas (λ). Como o sistema de informação do hospital registra apenas o número de admissões e o número de admissões recusadas é geralmente desconhecido, temos que aproximar (λ). Isso pode ser feito usando as expressões 3.3 e 3.4. Primeiro, o numerador em 3.3 é igual ao número médio de leitos ocupados, e consequentemente, após substituir P_c usando 3.3 teremos:

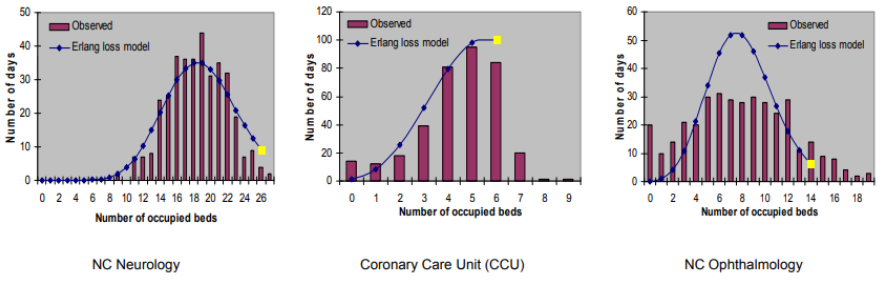
$$\text{Número médio de camas ocupadas} = \left(1 - \frac{\frac{(\lambda\mu)^c}{c!}}{\sum_{k=0}^{c}\frac{(\lambda\mu)^k}{k!}}\right)\lambda \cdot \mu \tag{3.5}$$

O número médio de leitos ocupados pode ser obtido no sistema de informações do hospital. Além disso, o número de leitos (c) e o ALOS (μ) são variáveis conhecidas. Após a substituição em 3.5 λ é o único parâmetro deixado desconhecido. Finalmente, o número de chegadas é determinado numericamente para todas as enfermarias clínicas.

Validação do modelo

Um dos principais objetivos deste estudo é aproximar o número de leitos necessários em uma enfermaria clínica. Portanto, primeiro determina-se a distribuição do número de leitos ocupados. Na 3.4 temos o número observado de leitos ocupados (às 08h00) comparado com o modelo de perda de Erlang, por exemplo, para três enfermarias. A curva do modelo de perda foi reduzida no número de leitos operacionais que era respectivamente 26 (Neurologia), 6 (UCC) e 14 (Oftalmologia).

Figura 3.4: O número de leitos ocupados observados versus o modelo de perda de Erlang



Percebe-se em 3.4 que em alguns dias o número de leitos ocupados excedem o número de leitos operacionais. Isso significa que existe uma alta demanda por parte dos pacientes, implicando que um eventual gerente de ala pode vir a decidir abrir temporariamente alguns leitos extras. Isso implica em uma consequente pressão no conjunto de funcionários que serão responsáveis por realocar tais recursos.

Além disso, o modelo parece razoável para o NC Neurology e o CCU, mas é razoavelmente pobre para a Oftalmologia de NC.

Para quantificar o qualidade do ajuste, introduzimos uma medida de validação. Primeiro definimos:

1. P_i = probabilidade de i leitos serem ocupados dentro do 'Erlang loss model'.

2. $Preal_i$= probabilidade de i leitos serem efetivamente ocupados na realidade.

3. $D_i = Preal_i$ - P_i para $i = 0, 1, ..., c-1$.

4. $D_c = \sum_{k=c}^{\infty} p_{real_k} - P_c$

A fórmula final acima é usada para comparar as possíveis situações diferentes com uma enfermaria totalmente ocupada. Lembre-se que na prática o número de leitos ocupados pode ocasionalmente exceder o número de leitos operacionais, **o que não é possível no modelo de perda de Erlang.** Para comparar até que ponto a distribuição empírica e a distribuição de ocupação de leitos dada pelo modelo de perda de Erlang são semelhantes, definimos nossa medida de desempenho para a qualidade de ajuste como a soma das diferenças absolutas entre as duas probabilidades:

$$\text{Qualidade de Ajuste} = 1 - \frac{1}{2}\sum_{i=0}^{c} \|D_i\| \tag{3.6}$$

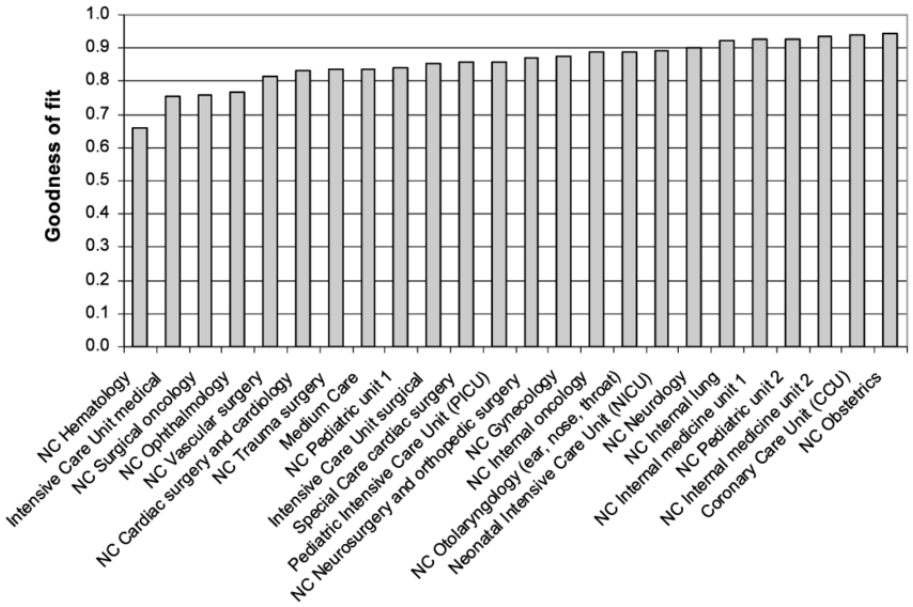
Em termos de funções de distribuição, nossa medida pode ser interpretada como a quantidade de densidade de probabilidade que a distribuição empírica e a distribuição de perda de Erlang têm em comum. Portanto, a medida é um número entre 0 e 1, onde 0 indica um ajuste muito ruim (sem massa de probabilidade em comum) e 1 significa que as probabilidades são iguais para todo o número de leitos ocupados (correspondência exata entre as funções de distribuição).

Tabela 3.2: Qualidade de ajuste do "Erlang loss model"ao descrever o número de leitos ocupados pelos pacientes

Ward Description	Goodness of fit	Ward Description	Goodness of fit
Coronary Care Unit (CCU)	0.941	NC Internal medicine unit 2	0.935
Intensive Care Unit surgical	0.854	NC Pediatric unit 1	0.841
Intensive Care Unit medical	0.756	NC Pediatric unit 2	0.926
Pediatric Intensive Care Unit (PICU)	0.860	NC Otolaryngology (ear, nose, throat)	0.889
Neonatal Intensive Care Unit (NICU)	0.891	NC Internal lung	0.921
Medium Care	0.838	NC Neurosurgery and orthopedic surgery	0.872
Special Care cardiac surgery	0.856	NC Neurology	0.902
NC Cardiac surgery and cardiology	0.833	NC Obstetrics	0.945
NC Gynaecology	0.874	NC Internal oncology	0.888
NC Hematology	0.658	NC Ophthalmology	0.768
NC Surgical oncology	0.759	NC Trauma surgery	0.835
NC Internal medicine unit 1	0.925	NC Vascular surgery	0.815

Obteve-se uma média de Qualidade de Ajuste de 0.86 significando que **o modelo empírico e as distribuições de ocupação de leito baseadas no modelo de Erlang têm em média 86% da massa de probabilidade em comum.**

Figura 3.5: Qualidade de ajuste do "Erlang loss model"para as 24 clinicas



Outra observação importante e mais geral é que a variabilidade tanto no número de chegadas por dia e do LOS resulta em grandes flutuações de carga de trabalho, ou seja, uma variabilidade considerável na ocupação do leito. Portanto, a ocorrência de super e sublotação é inevitável tornando a flexibilidade no planejamento da força de trabalho em enfermarias clínicas crucial.

Avaliando o tamanho das enfermarias clínicas

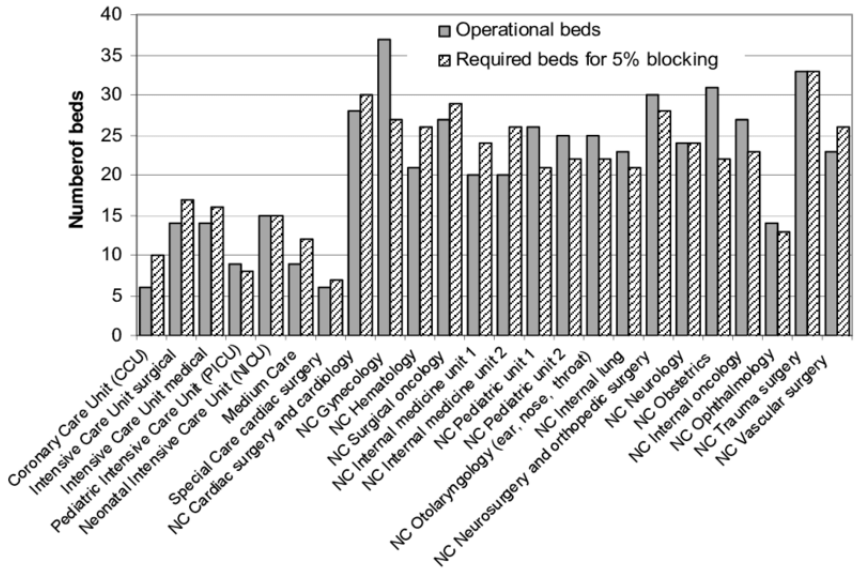
Primeiramente, avalia-se o tamanho atual das enfermarias do hospital. Para cada enfermaria, o número de chegadas é determinado conforme descrito na 3. Em seguida, o número de leitos necessários é calculado para três níveis de bloqueio (2, 5 e 10%). Consulte a 3.3 para obter os resultados.

Tabela 3.3: Número de leitos requeridos para diferentes níveis de super-lotação na situação recorrente (2006)

Ward description	Operational beds (2006)	Number of daily arrivals (λ)	Number of beds required for:		
			2% Blocking	5% Blocking	10% Blocking
Coronary Care Unit	6	3.52	12	10	9
Intensive Care Unit surgical	14	2.26	19	17	15
Intensive Care Unit medical	14	2.14	18	16	14
Pediatric Intensive Care Unit	9	0.97	9	8	7
Neonatal Intensive Care Unit	15	1.36	17	15	14
Medium Care	9	3.06	13	12	10
Special Care cardiac surgery	6	1.79	8	7	6
NC Cardiac surgery and cardiology	28	5.62	33	30	27
NC Gynecology	37	6.84	30	27	24
NC Hematology	21	7.57	29	26	24
NC Surgical oncology	27	3.55	32	29	26
NC Internal medicine unit 1	20	2.9	27	24	21
NC Internal medicine unit 2	20	4.17	29	26	23
NC Pediatric unit 1	26	4.5	23	21	18
NC Pediatric unit 2	25	4.02	24	22	20
NC Otolaryngology (ENT)	25	3.8	24	22	19
NC Internal lung	23	3.32	23	21	18
NC Neuro- and orthopedic surgery	30	4.26	31	28	25
NC Neurology	24	3.29	26	24	21
NC Obstetrics	31	11.14	25	22	20
NC Internal oncology	27	3.86	25	23	20
NC Ophthalmology	14	5.18	14	13	11
NC Trauma surgery	33	3.97	36	33	30
NC Vascular surgery	23	3.19	29	26	23
Total	**507**	**95.98**	**556**	**502**	**445**

Podemos inferir acerca de qual porcentagem de bloqueio (ou admissões recusadas) é razoável e, portanto, aceitável é sujeito a discussão. Os formuladores de políticas hospitalares geralmente se referem a uma meta de 5%, mas a consequência de tal escolha em termos de requisitos de capacidade muitas vezes não é reconhecida. Tabela 3.3 revela que, para uma meta de 5%, o número total de leitos necessários em 24 enfermarias é aproximadamente 500, o que não está muito longe do número de leitos operacionais em 2006, que é de 507. Na Figura 3.6, o número de leitos operacionais é comparado com o número de leitos necessários para cada enfermaria (5% bloqueio).

Figura 3.6: O número de leitos solicitados versus o número de leitos operacionais

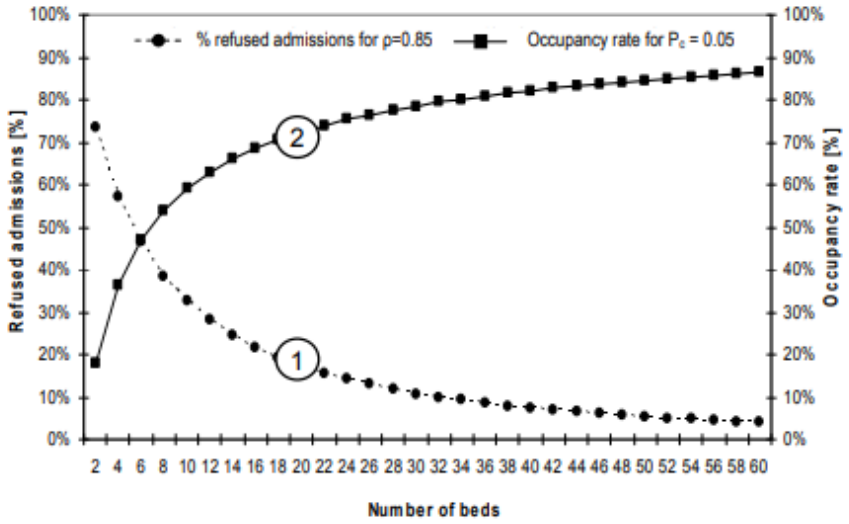


A razão (leitos necessários/leitos operacionais) varia de 0,71 a 1,67. Em outras palavras, algumas enfermarias têm muitos leitos, enquanto outras têm uma grave escassez.

O impacto da fusão de departamentos na eficiência do hospital

Na maioria dos sistemas de filas, ocorrem economias de escala, o que significa que sistemas de atendimento maiores podem operar com taxas de ocupação mais altas do que as menores, ao mesmo tempo em que atingem a mesma porcentagem de bloqueio ou atraso (WHITT, 1992). A figura 3.7 ilustra o impacto dramático para a situação na maioria dos Hospitais holandeses onde os tamanhos das enfermarias são relativamente pequenos e dispersos e onde 85% da taxa de ocupação alvo tornou-se um padrão de ouro. Na Figura 3.7, dois gráficos são mostrados para vários tamanhos de ala ($2 \leq c \leq 60$):

Figura 3.7: Relação entre o número de leitos, fração de admissões recusadas (P_c) e taxas de ocupação(ρ) nos hospitais holandeses



Se as economias de escala forem aplicadas adequadamente, efetivando a fusão de departamentos (leito compartilhado) ou misturando os fluxos de pacientes, ambos em um nível de serviço aceitável (em termos de admissões recusadas) uma taxa de ocupação economicamente viável pode ser alcançada.

Estudo de Caso

Podemos exemplificar um cenário em que o DSS pode ser usado para explorar o benefício potencial da fusão de departamentos. Fazemos isso virtualmente fundindo a 'coronary care unit' (CCU), 'medium care' (MC) e 'special care cardiac surgery' (SC), pois o nível e tipo de cuidado são semelhantes. Ao selecionar diferentes unidades, o usuário pode experimentar com facilidade e rapidez com diferentes cenários no DSS. A Tabela 3.4 resume as estatísticas de 2006. Os resultados do cenário onde essas três unidades são mescladas também é apresentado nesta Tabela.

Tabela 3.4: O efeito da fusão de departamentos na eficiência operacional: estudo de caso

Parameters (2006)	**CCU**	**MC**	**SC**
Operational beds	6	9	6
ALOS [days]	1.694	2.374	1.715
Occupancy rate	73.3%	69.8%	48.2%
Fraction of refused admissions*	26.2%	13.5%	5.61%
Number of beds required for 5% blocking*	10	12	7
After merging CCU, MC, and SC			
ALOS (weighted)		1.96	
Number of beds required for 5% blocking*		**22**	
Occupancy rate		**71.7%**	

*Calculated with the Erlang loss model

Na situação recorrente em que o estudo foi realizado essas três unidades somavam 21 leitos e

o número de admissões recusadas variava de 5,61% (SC) a 26,2% para o UCC. O número total de camas necessárias,para 5% de admissões negadas em cada enfermaria, é 29.

Após a fusão, o número total de leitos necessários para esse mesmo nível de serviço (bloqueio de 5%) é 22, apenas um leito a mais do que o número de leitos neste momento. Assim, a melhora na eficiência operacional é significativa e pode ser alcançada apenas criando uma escala maior.

Deep Generative Models for Queuing Systems

Dada uma série de N chegadas associadas aos seus tempos de partida $\{a_i, d_i\}_{i=1}^{N}$, cada par representado em tempo contínuo $\mathbb{R}^+$, considere o sistema de filas G/G/∞ para o qual os tempos de espera w_i são nulos e os tempos de atendimento são dados por $s_i = d_i a_i$.Cada chegada possui um conjunto de co-variáveis $\{x_i \epsilon \mathbb{R}^+\}_{i=1}^{N}$ a eles associados (como, por exemplo, os locais de embarque e desembarque para corridas de táxi das quais a duração do trajeto depende fortemente). Assumimos que nossos tempos de serviço são amostrados de uma distribuição desconhecida $P_D\left(S|x_i, H_i\right)$ condicionados com a co-variáveis e o histórico de chegadas: $H \equiv \{a_1, ..., ai\}$. O objetivo é aprender um modelo generativo com distribuição de probabilidade implícita P_θ que tem por objetivo aproximar P_D. Em poucas palavras, a abordagem a ser tomada consiste em modelar os tempos de chegada e atendimento do cliente separadamente:

- **Modelagem de Tempos de chegada** : Primeiramente, se modela a sequencia de tempos de chegada $\{a_i\}_{i=1}^{N}$ Utilizando o modelo 3 , satisfazendo tanto a escalabilidade quanto a robustez no que diz respeito "modelling misspecification"

- **Modelagem de Tempo de servidores** : Utiliza-se os "hidden states"(estados omitidos) do modelo RPP(Recurrent Point Process) treinado codificando os H_i, juntamente com as co-variantes x_i, como input do modelo de tempo de servidores : o Recurrent Adversarial Service Times (RAS). O modelo RAS mapeia conjuntamente os inputs de forma determinística com vetores aleatórios de um espaço latente para o espaço das observações, e é parametrizado por uma segunda RNN(Recurrent Neural Network), que é caracterizada por conseguir capturar a característica da dinâmica que molda o sistema de servidores recorrente.

Temos que o modelo do tipo 'Deep Generative Service Times' é hierárquico e trata os sistemas de servidores como uma única fila do tipo G/G/∞ , ao passo que não se necessita de fazer pressupostos acerca da disciplina da fila.

Recurrent Point Process (RPP)

Dentro da teoria das filas, as chegadas dos clientes definem um Processo de Ponto unidimensional. Definimos $f^*\left(t\right)$ como a função de verossimilhança de um ponto unidimensional induzido por uma função densidade $\lambda^*\left(t\right)$ onde a $\lambda^*\left(t\right)$ é definida por uma RNN(Recurrent Neural Network). Consideramos uma ponto unidimensional com o suporte compacto $S \subset \mathbb{R}$. Formalmente, a função de verossimilhança será escrita por intermédio de um processo de Poisson não homogêneo entre

chegadas, condicionado ao histórico de chegadas H_i. Para um processo unidimensional a função de verossimilhança que condiciona a existência da próxima chegada no tempo t será:

$$f^*(t) = \lambda^*(t)\, exp\left\{\int_{a_i}^{t} \lambda^*\left(t'\right) dt'\right\} \tag{3.7}$$

onde a função de intensidade λ^* é uma função (localmente) integrável. A dependência funcional da função de intensidade é dada por uma RNN(Recurrent Neural Network) com o hidden state $h_i^a \epsilon \mathbb{R}^h$, onde uma função exponencial garante que a intensidade é não negativa:

$$\lambda^*(t) = exp\left\{v^t \cdot h_i^a + w^t\,(t - a_i) + b^t\right\} \tag{3.8}$$

Onde o vetor $v^t \epsilon \mathbb{R}^h$ e as escalares w^t e b^t são as variáveis treináveis . A equação atualizada para as "hidden variables"da RNN(Recurrent Neural Network) pode ser escrita com o auxilio de uma função não-linear:

$$h_i^a = g_\theta\left(a_i, h_{i-1}^a\right) \tag{3.9}$$

RPP Loss - Substituindo a equação 3.9 na 3.8 e integrando em função do tempo obtemos f^* como uma função de h_i^a. Os parametros do modelo serão obtidos maximizando a função log-verossimilhança conjunta do modelo:

$$\iota_{RR} = \sum_{i=1}^{N} log f^*\left(\delta_{i+1} | h_i^a\right) \tag{3.10}$$

Onde $\delta_{i+1} = a_{i+1} - a_i$ denota o tempo entre as chegadas (inter-arrival time) e N é o número total de chegadas observadas.

Recurrent Adversarial Service Times (RAS)

Para modelar distribuições gerais de tempo de serviço, e evitar a necessidade de especificar qualquer disciplina de serviço, consideramos um modelo generativo definido pelo seguinte processo de duas etapas:

1. Amostrar um vetor aleatório $Z \epsilon \mathbb{R}^D$ de uma distribuição conhecida e fácil de amostrar $P(Z)$.

2. Aloque z juntamente com a representação em RPP(Recurrent Point Process) via h_i^a e as co variáveis x_i da i-ésima chegada, através de uma função paramétrica $\Phi_\theta : \mathbb{R}^D \times \mathbb{R}^h \times \mathbb{R}^c \to \mathbb{R}^+$ com conjunto de parâmetros θ, que gera amostras de tempo de serviço seguindo uma distribuição P_θ. Vamos mover essa distribuição para “próximo” da distribuição empírica de tempo de serviço P_D variando θ para minimizar uma distancia razoável entre ambas as distribuições.

Gerador de tempo de serviço

Φ_θ- Para cada tempo de chegada a_i, primeiramente amostramos uma variável aleatória normalmente distribuída $z_i \sim N(0,1)$ e posteriormente definimos as "hidden representations"$u_i \in \mathbb{R}^D$ tendo assim a "rectified linear unit"(ReLU):

$$u_i = ReLU(W_a^u h_i^a + W_x^u x_i + W_z^u z_i + b^u) \tag{3.11}$$

Onde $(W_a^u$, $(W_x^u$, $(W_z^u$ e $b^u \in \theta$ são parâmetros treináveis. Tal representação omite a informação tanto da covariáveis x_i quanto histórico de chegada H_i através de h_i^a definido em 3.9. Por conseguinte, modelamos a dinâmica de resposta do sistema para clientes recém-chegados com uma RNN(Recurrent Neural Network) cujo estado oculto $h_i^\Phi \subset \mathbb{R}^D$ definida:

$$h_i^\Phi = g_\theta\left(u_i h_{i-1}^\Phi\right) \tag{3.12}$$

Por fim, computamos o tempo de serviço via:

$$s_i = exp(W_h^s h_i^\Phi + W_z^s z_i + b^s) \tag{3.13}$$

onde $z_i \sim N(0,1)$ e W_h^s, W_z^s e $b^s \subset \theta$ são parametros treináveis, e a exponencial é escolhida para restringir as amostras de $\mathbb{R}^+$. A composição das funções 3.11, 3.12 e 3.13 definirão o **Gerador dado por** $\Phi_\theta = \Phi_\theta(z, h_i^a, x_i)$. Vale ressaltar que se acrescentou uma "noise source"em 3.13 para aumentar a variabilidade das amostras.

Para treinar o modelo opta-se por escolher o método **Wasserstein Loss and Training** a distancia Wassertein-1 que viabiliza o cálculo entre as distâncias de duas distribuições P_D e P_θ sendo definida como:

$$W(P_\mathcal{D}, P_\theta) = \inf_{\gamma \in \Pi(P_\mathcal{D}, P_\theta)} \mathbb{E}_{(x,y)\sim\gamma}\left[||x-y||\right] \tag{3.14}$$

Onde o termo $\prod(P_D, P_\theta)$ caracteriza o conjunto de todas as distribuições conjuntas $\gamma(x,y)$ cujas marginais são, respectivamente P_D e P_θ. Como o mínimo acima é, em geral, intratável, retorna-se para a Dualidade de Kantorovich-Rubinstein que certifica que:

$$W(P_\mathcal{D}, P_\theta) = \sup_{f \in \mathfrak{L}_1} \mathbb{E}_{s\sim P_\mathcal{D}}\left[f(s)\right] - \mathbb{E}_{s\sim P_\theta}\left[f(s)\right] \tag{3.15}$$

Onde o supremo está sobre todas as funções de 1-Lipschitz, e é ainda (em geral) intratável. O truque é então restringir o espaço de busca ao de uma família parametrizada de funções $f\varphi$, a função crítica, que se modela com redes neurais e aprende sob a restrição de 1-Lipschitz. Agora, nosso objetivo é aprender distribuições não apenas condicionadas a um conjunto de covariáveis, mas também ao processo estocástico de chegada. Isso é por que não podemos usar diretamente, de forma direta, o bem solução ótima conhecida da 3.14 em uma dimensão. Em vez disso, definimos nosso crítico como função de ambos h_i^a e x_i, sendo assim $f\varphi = f\varphi(s, h_i^a, x_i)$ de modo a

encontrarmos de fato a distancia para cada condicional. Podemos assim escrever a "RAS loss function":

$$\begin{aligned}\mathcal{L}(\theta) = \max_{\varphi} \mathbb{E}_{(a_i,\mathbf{x}_i)\sim P_{\mathcal{D}}} \Big\{ & \mathbb{E}_{s_i\sim P_{\mathcal{D}}}\left[f_\varphi(s_i,\mathbf{h}_i^a,\mathbf{x}_i)\right] \\ & - \mathbb{E}_{s_i\sim P_\theta}\left[f_\varphi(s_i,\mathbf{h}_i^a,\mathbf{x}_i)\right] \\ - \mathbb{E}_{s_i\sim P^*}\left[\left(\max\left\{0, |\nabla_{s_i} f_\varphi(s_i,\mathbf{h}_i^a,\mathbf{x}_i)| - 1\right\}\right)^2\right] & \Big\}\end{aligned} \tag{3.16}$$

Logo, percebe-se que o último termo na equação 3.16 impõe aproximadamente a restrição de 1-Lipschitz forçando a norma do gradiente de f_φ para ser no máximo 1 ao longo das linhas retas amostradas dos espaços probabilísticos P^*. Logo, ao minimizar a função 3.16 com respeito a θ sob uma função crítica ótima que minimiza a distancia de Wasserstein-1 entre P_D e P_θ,isso define o "jogo contraditório"podemos assim sumarizar esse algoritmo:

Algorithm 1: Recurrent Adversarial Service Time

1: **Requires:** Dataset $\mathcal{D} = \{(a_i, s_i, \mathbf{x}_i)\}_{i=1}^N$, Critic Iterations Number n_c and the penalty weight λ.
2: **while** θ not converged **do**
3: **for** $i = 1, ..., N$ **do**
4: Draw $\{(a_i, s_i, \mathbf{x}_i)\} \sim P_{\mathcal{D}}$
5: **for** $k = 1, ..., n_c$ **do**
6: Draw $\mathbf{z} \sim P(Z)$ and $\delta \sim \text{Uniform}(0, 1)$,
7: $\mathbf{h}_i^a \leftarrow g_\rho(a_i, \mathbf{h}_{i-1}^a)$,
8: $\tilde{s}_i \leftarrow \Phi_\theta(\mathbf{z}, \mathbf{h}_i^a, \mathbf{x}_i)$,
9: $\hat{s}_i \leftarrow \delta s_i + (1-\delta)\tilde{s}_i$,
10: $\mathcal{L}_{\theta,\varphi} \leftarrow f_\varphi(\tilde{s}_i, \mathbf{h}_i^a, \mathbf{x}_i) - f_\varphi(s_i, \mathbf{h}_i^a, \mathbf{x}_i)$
11: $+\lambda \left(\max\left\{0, |\nabla_{\hat{s}_i} f_\varphi(\hat{s}_i, \mathbf{h}_i^a, \mathbf{x}_i)| - 1\right\}\right)^2$,
12: $\varphi \leftarrow \text{Adam}(\nabla_\varphi \mathcal{L}_{\theta,\varphi})$
13: **end for**
14: Draw $\mathbf{z} \sim P(Z)$
15: $\theta \leftarrow \text{Adam}(\nabla_\theta(-f_\varphi(\Phi_\theta(\mathbf{z}, \mathbf{h}_i^a, \mathbf{x}_i))))$
16: **end for**
17: **end while**
18: **return** Φ_θ

Perceba que esse algoritmo sumariza o modelo RAS e vale ressaltar que a linha 7 trata os parâmetros do modelo RPP como fixos e optimizáveis.

Planejando o Campo experimental

Quando a estrutura experimental, opta-se por postular conjuntos de dados sintéticos com modelos bem estabelecidos para chegadas e processos de atendimento, bem como conjuntos de dados empíricos .Modelar esses datasets demonstra a capacidade de nossa abordagem para lidar com diversas áreas de aplicação de forma flexível e escalável. **Synthetic Datasets**- Com o intuito de fornecer um ambiente controlado para testar o comportamento dessa metodologia, seleciona-se os seguintes conjuntos de dados para diferentes chegadas e serviços processos:

1. Processos de chegada: São considerados dois processos distinto, sendo (i) o Processo de Hawkes (H), que é um modelo para fenômenos auto-estimulantes onde as chegadas de usuários aumenta a probabilidade de outros usuários chegarem, e que é definido por meio da função de intensidade condicional:

$$\lambda_H(t) = \lambda^* + e^{-\beta t}(\lambda_0 - \lambda^*) + \sum_{a_i : t > a_i} \alpha\mu(t - a_i) \tag{3.17}$$

onde λ_0 e λ^* são os valores iniciais da intensidade do processo para eventos exógenos (eventos chegada),α, β são seus parâmetros de salto e decaimento, e o kernel de memória $\mu(t) = e^{-\beta t}$ fornece a intensidade dada pelas chegadas passadas (a_i) ; (ii) o Processo de Hawkes não linear (NH), que é uma extensão do processo de Hawkes que permite o comportamento inibitório através de uma função não linear sobre o histórico de chegadas.

2. Modelos de serviço: introduzimos duas disciplinas do sistema de serviço: (i) a distribuição de Phase-Type (PT), em que se decompõe o serviço como uma série de etapas exponenciais de serviço, e que é definido com o tempo decorrido entre o estado inicial e o de absorção e uma cadeia de markov continua; (ii) a distribuição Processor Sharing (PS), um modelo de filas no qual o sistema lida com uma quantidade infinita de clientes simultaneamente, mas deve realocar recursos com cada chegada ou partida de um novo cliente. Pode-se pensar que cada cliente no sistema recebe instantaneamente $\frac{1}{Q(t)}$ do poder do servidor, em qualquer tempo, onde $Q(t)$ é o tamanho da fila. As combinações dos dois tipos de chegada e dos dois de modelos de servidores produz quatro conjuntos de dados sintéticos diferentes, rotulados como: H-PS, NH-PT e NH-PS. Simulou-se 5×10^6 datapoins para os modelos de Hawkes, e 5×10^5 datapoints para suas extensões não lineares.

Empirical Datasets- Reuniram-se conjuntos de dados de uma variedade de serviços de internet, para fornecer uma análise aprofundada acerca dos padrões temporais de usuários em diferentes domínios.

1. Stackoverflow: uma plataforma de perguntas e respostas para programadores. Definimos as chegadas de clientes como os pontos em que as perguntas são postadas pelos usuários da página da web, e o tempo de serviço como o tempo decorrido entre uma pergunta e sua subsequente resposta aceita. Como co-variáveis, tomam-se as cinco primeiras 'tags' de cada pergunta. Essa visão estabelece o conjunto de usuários que fornecem respostas como o sistema de serviço. Analisando um total de 2×10^7 questões.

2. Github: O repositório de controle de versão e Internet serviço de hospedagem. Defini-se a criação de um problema em um determinado repositório conforme as chegadas dos clientes. Os horários de saída são os momentos em que os problemas dados são fechados. Não houve escolha de co-variáveis para este conjunto de dados. O conjunto de usuários associados a um determinado repositório pode ser pensado como o próprio servidor do sistema. Analisaram-se os 500 principais (classificados pelo número de edições) repositórios na plataforma em 2015, para um total de $1,5 \times 10^6$ questões diferentes.

3. Conjunto de dados de táxi da cidade de Nova York (NY): o conjunto de dados contém dados de viagens de táxi individuais na cidade de Nova York. As chegadas dos clientes são definidas como a hora de início da viagem e a hora de partida como a hora de fim da viagem. Como co-variáveis foram escolhidos os pontos inicial e final das viagens. Aqui o sistema de atendimento é dado tanto pelo táxi que presta o serviço como pelo transporte dados pelas possíveis ruas e rodovias pertencentes à cidade de Nova York. Analisando cerca de $1,1 \times 10^7$ viagens.

	NH-PT		NH-PS		H-PS		Github		NY		Stackoverflow	
mean ($\bar{s}$)	0.052		0.0004		2.125e-5		0.0113		0.0068		0.0193	
	error	KS	error	KS	error	KS	error	KS	error	KS	error	KS
NTE	0.410	0.289	0.0450	0.410	2.44e-2	0.766	0.071	0.388	0.062	0.321	0.380	0.502
NS	0.209	0.154	0.0006	0.082	**2.18e-5**	0.401	0.096	0.341	0.025	0.154	0.378	0.466
ATE	0.219	0.193	0.0371	0.870	3.96e-3	0.199	0.217	0.517	0.078	0.126	0.382	0.233
AS	0.215	0.113	0.0016	0.448	1.24e-4	0.121	**0.071**	0.039	0.006	0.094	0.383	**0.226**
RATE	0.218	0.124	0.0031	0.062	1.37e-4	0.136	0.112	0.240	0.098	0.165	0.388	0.492
RAS	**0.207**	**0.094**	**0.0005**	**0.042**	1.09e-4	**0.110**	0.072	**0.034**	**0.005**	**0.030**	**0.369**	0.281

Figura 3.8: Avaliação dos modelos em todos os conjuntos de dados (em unidades arb.). KS - Teste de Kolmogorov-Smirnov.

O erro de predição considerado foi definido como $\frac{1}{N}\sum_i |s_i - \langle \tilde{s}_i \rangle|$, onde N denota o tamanha do dataset , s_i seus valores empíricos e $\langle \tilde{s}_i \rangle$ sua predição média obtida via amostragem de monte carlo. O Kolmogorov-Smirnov (KS) foi calculado entre as distribuições empíricas e as distribuições geradas. Da tabela podemos perceber que o RAS é de fato o mais efetivo.

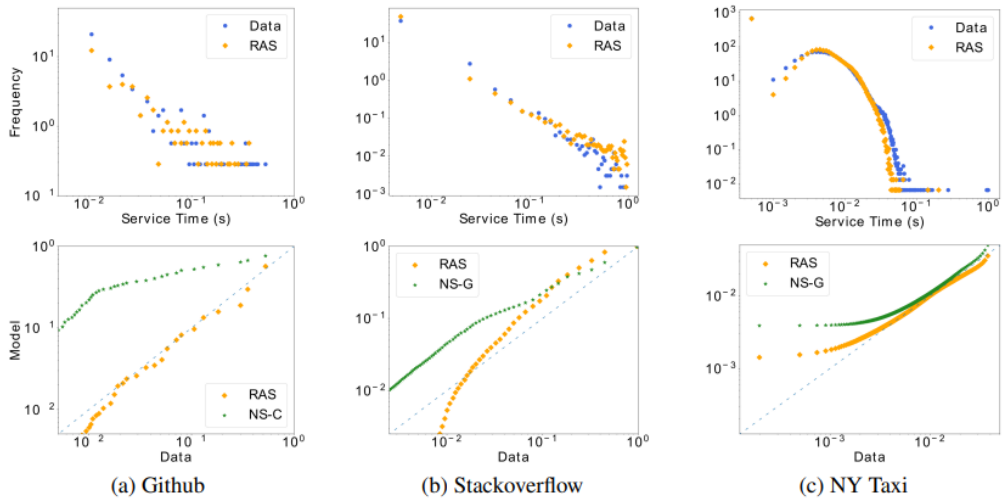


Figura 3.9: (Topo)Comparação entre as distribuições empíricas de tempos de atendimento (test set) e dados gerados pelo RAS. (Inferior) Q-Q gráficos contra distribuições empíricas para o melhor modelo NS e modelo RAS.

A força do RAS é ainda mais aparente em suas formas de distribuição. A 3.9 mostra histogramas de dados gerados com RAS contra a distribuição de dados empíricos, bem como Q-Q plots do melhor modelo NS e modelo RAS versus os dados. Gráficos semelhantes para os conjuntos de dados sintéticos podem ser encontrados no Material Suplementar. Note que os prazos em esses conjuntos de dados abrangem até quatro ordens de grandeza. O modelo RAS fornece

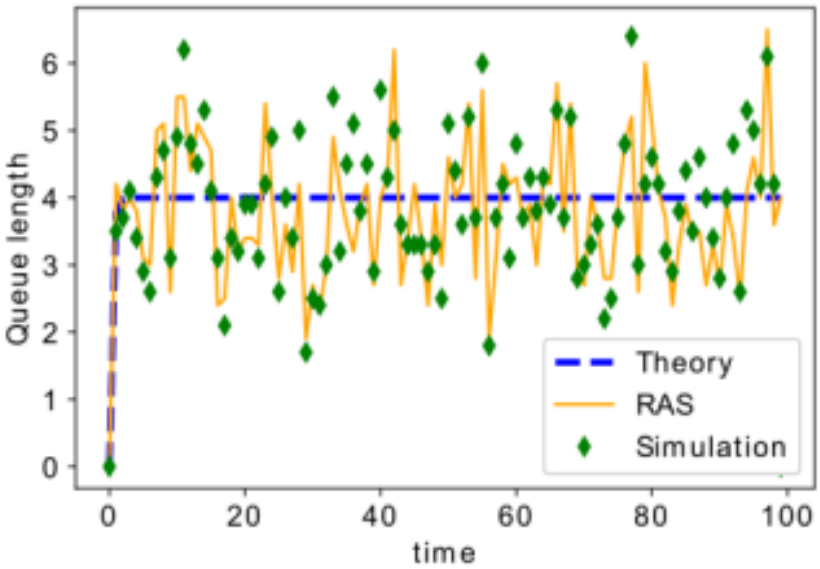
um ajuste muito melhor aos dados do que os modelos NS (ver gráficos Q-Q). Em particular, RAS captura exclusivamente o comportamento de longo e curto prazo. O último é aparente no canto superior esquerdo dos gráficos do histograma. O O modelo NS (Neural Service), por outro lado, fornece apenas o comportamento correto da cauda longa, pois é limitado pelas formas distributivas dos outputs.

Por fim, devemos comparar os resultados com um referencial teórico: **Comparison with Theory**- Vamos agora considerar novamente a fila do tipo Hawkes/Phase-Type/∞ para a qual o comprimento médio da fila que depende do tempo mostrou ser :

$$\langle Q(t)\rangle = \lambda_\infty \left(-S^T\right)^{-1}\left(1-e^{S^T}\right)w - (\lambda_0-\lambda_\infty)\times\left(S^T+(\beta-\alpha)\,\mathbb{I}\right)^{-1}\left(\mathbb{I}e^{-(\beta-\alpha)t}-e^{S^T t}\right)w \tag{3.18}$$

Onde $_0$, α e β são parâmetros do processo de Hawkes (veja em 3.17) e $\lambda_\infty = \dfrac{\beta\lambda^*}{\beta-\alpha}$ é valor estacionário da sua intensidade. Simulou-se uma fila Hawkes/Phase-Type/∞ e treinou-se RAS no dataset resultante. Podemos assim estimar $\langle Q(t)\rangle$ usando RAS amostrando um tempo de serviço para cada chegada no train test, contando e rastreando o número de clientes em serviço e tomando a média sobre o conjunto de teste. Agora, comparando com supracitado referencial teórico teremos:

Figura 3.10: Comprimento médio da fila tempo-dependente $\langle Q(t)\rangle$ para a fila Hawkes/PhaseType/∞ fila (H-PT).



Temos assim que a figura 3.10 mostra os resultados em comparação com o momento exato (dado em 3.18), e o test set simulado. Temos assim que se foi apresentada uma solução não-paramétrica para o aprendizado de distribuição de tempo de atendimento em filas. Sistemas com distribuições de chegada generalizadas. Tal metodologia superou todas as linhas de base e reproduziu distribuições de tempo de serviço complexas , inferindo acerca de características tanto multimodais quanto de caudas longas.

Referências Bibliográficas

BEKKER, R.; BRUIN, A. M. de. Time-dependent analysis for refused admissions in clinical wards. *Annals of Operations Research*, Springer, v. 178, p. 45–65, 2010.

KENDALL, D. G. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded markov chain. *The Annals of Mathematical Statistics*, JSTOR, p. 338–354, 1953.

MITRANI, I. Computer system models. *Handbooks in Operations Research and Management Science*, Elsevier, v. 3, p. 519–559, 1992.

NEWELL, C. *Applications of queueing theory*. [S.l.]: Springer Science & Business Media, 2013. v. 4.

UTLEY, M. et al. Analytical methods for calculating the distribution of the occupancy of each state within a multi-state flow system. *Ima journal of management mathematics*, Oxford University Press, v. 20, n. 4, p. 345–355, 2009.

UTLEY, M. et al. Analytical methods for calculating the capacity required to operate an effective booked admissions policy for elective inpatient services. *Health care management science*, Springer, v. 6, p. 97–104, 2003.

WHITT, W. Understanding the efficiency of multi-server service systems. *Management Science*, INFORMS, v. 38, n. 5, p. 708–723, 1992.

WHITT, W. Queues with time-varying arrival rates: A bibliography. *Available on http://www. columbia. edu/˜ ww2040/TV_ bibliography_ 091016. pdf*, 2016.

WORTHINGTON, D. Reflections on queue modelling from the last 50 years. *Journal of the Operational Research Society*, Springer, v. 60, p. S83–S92, 2009.