

ALINIE MENDES FIRMINO

**A IMPORTÂNCIA DA GESTÃO DE FILAS NA PRESTAÇÃO DE
SERVIÇOS: UM ESTUDO DE CASO EM UMA EMPRESA DE
*OUTSOURCING***

São Paulo
2014

ALINIE MENDES FIRMINO

**A IMPORTÂNCIA DA GESTÃO DE FILAS NA PRESTAÇÃO DE
SERVIÇOS: UM ESTUDO DE CASO EM UMA EMPRESA DE
*OUTSOURCING***

Monografia apresentada à Escola
Politécnica da Universidade de São Paulo
para obtenção do título de Especialista em
Gestão e Engenharia da Qualidade

Orientador: Prof. Dr. Adherbal Caminada
Netto

São Paulo
2014

ALINIE MENDES FIRMINO

**A IMPORTÂNCIA DA GESTÃO DE FILAS NA PRESTAÇÃO DE
SERVIÇOS: UM ESTUDO DE CASO EM UMA EMPRESA DE
*OUTSOURCING***

Monografia apresentada à Escola
Politécnica da Universidade de São Paulo
para obtenção do título de Especialista em
Gestão e Engenharia da Qualidade

Orientador: Prof. Dr. Adherbal Caminada
Netto

São Paulo
2014

MBAEQ
F5182

DEDALUS - Acervo - EPMN



31600022731

FICHA CATALOGRÁFICA

Firmino, Alinie Mendes

A importância da gestão de filas na prestação de serviços: um estudo de caso em uma empresa de Outsourcing / A.M. Firmino.
— São Paulo, 2014.

p.59

Monografia (MBA em Gestão e Engenharia da Qualidade)
Escola Politécnica da Universidade de São Paulo. Programa de
Educação Continuada em Engenharia.

1. Tecnologia da informação. 2. Gestão de filas I. Universidade de
São Paulo. Escola Politécnica. Programa de Educação Conti-
nuada em Engenharia II.t.

Ao meu marido, Marcos Vinícius B. Peigo, que sempre me incentivou para a realização dos meus ideais, encorajando-me a enfrentar todos os momentos difíceis da vida.

Aos meus pais, Rejane Cristina Mendes e Gilmar Pacheco Firmino, pela compreensão, apoio e contribuição na minha formação acadêmica.

AGRADECIMENTOS

Ao meu marido, Marcos Vinícius B. Peigo, que esteve ao meu lado em todos os momentos, me dando força e me incentivando todos os dias. Sem ele, certamente não teria chegado ao final do curso.

Aos meus colegas de trabalho que, direta ou indiretamente, contribuíram com insumos para o desenvolvimento desse trabalho.

Ao meu orientador, Prof. Dr. Adherbal Caminada, pelas lições ensinadas ao longo do curso que certamente trouxeram reflexo em todas as vertentes da minha vida.

"A vida não dá nem empresta; não se comove nem se apieda. Tudo quanto ela faz é retribuir e transferir aquilo que nós lhe oferecemos" (Albert Einstein)

RESUMO

Com o aumento da competitividade, a busca pela racionalização de recursos a fim de maximizar o lucro das empresas de serviços de tecnologia tem induzido a implementação de uma série de mudanças na gestão da produção. Considerando como o principal meio de entrega de um serviço de tecnologia o uso de uma fila de atendimento, cabe analisar as entradas e tempos decorridos nesse fluxo como forma de organização e otimização da cadeia como um todo. É nesse contexto que a teoria das filas, amplamente utilizada em outros ramos, pode ser adaptada para a análise da capacidade produtiva de uma empresa de serviços, transformando-se em importante ferramenta de apoio para que o serviço entregue seja escalável, aumentando a produtividade, reduzindo os custos e aumentando os lucros. Ainda que a aplicabilidade da teoria possa ser comprometida, uma vez que o alto grau de variabilidade causado pelo cliente e a consequente despadronização impactam na coleta e análise adequada dos dados, com algum esforço de padronização é possível garantir o correto controle da produção. É fato que esses esforços são recompensados pela geração de benefícios visíveis para o cliente, que permitem não somente a manutenção dos contratos existentes, mas também a consolidação de diferenciais competitivos fundamentais, que garantem novos contratos e incrementam a rentabilidade da organização.

Palavras-chave: Teoria das filas. Controle da produção. Atendimento. *Outsourcing*.

ABSTRACT

With the increase of competitiveness and the quest for rationalizing resources in order to maximize profits, technology services companies have been facing a series of changes in production management. Considering that the main way of delivering services is through a service queue, we need to analyze the inputs and elapsed times within this flow as an opportunity for optimize the chain and the organization as a whole. It is in this context that the Queuing Theory, widely used in other industries, can be adapted to the analysis of the productive capacity of an IT services company becoming a tremendous support tool to assure that the delivered service is scalable, increasing its productivity, reducing its costs and raising profitability. The applicability of this theory could be compromised because the high degree of variability and therefore the lack of standardization caused by the customer, with serious impact in the proper collection and analysis of data. Even tough, some standardization effort can ensure proper control of production and still produce visible benefits to the customer, allowing not only the maintenance of existing contracts, but also increasing the companies win rate up on new contracts. Nevertheless, is a fact that the competitive advantages shall be strengthened as well.

Keywords: Queues Theory. Production Control. *Outsourcing*

LISTA DE ILUSTRAÇÕES

Figura 1	Variação mês/igual mês do ano anterior dos serviços de informação e comunicação, por Unidades da Federação	24
Figura 2	Ilustração de um sistema de filas	31
Figura 3	Fluxo de entrega de serviço	39

LISTA DE TABELAS

Tabela 1	Indicadores de receita nominal do setor de serviços, segundo grupos de atividades	23
----------	---	----

LISTA DE GRÁFICOS

Gráfico 1	Densidade de probabilidade	29
Gráfico 2	Evolução do <i>Backlog</i> em 12 meses. Situação atual. Demanda constante	54
Gráfico 3	Evolução do <i>Backlog</i> em 12 meses. Situação atual. Demanda variável	54
Gráfico 4	Evolução do <i>Backlog</i> em 12 meses. Situação proposta	55

LISTA DE QUADROS

Quadro 1	Quadro resumo de análise taxa de chegada e velocidade de atendimento	44
Quadro 2	Quadro resumo de análise do tempo médio em fila	46
Quadro 3	Quadro resumo de análise taxa de congestionamento	49
Quadro 4	Simulação final de equilíbrio do sistema	53

LISTA DE SÍMBOLOS

c	Número de atendentes do sistema
λ	Taxa média de chegada de clientes
μ	Taxa média de atendimento por atendente
ρ	Taxa de utilização do sistema
W_q	Tempo do cliente em fila
W_s	Tempo do cliente durante o atendimento
W	Tempo total de um cliente no sistema
L_q	Número de clientes na fila
L_s	Número de clientes em atendimento
L	Número total de clientes

SUMÁRIO

1	INTRODUÇÃO	16
1.1	Objetivo e escopo	17
2	METODOLOGIA	18
3	REVISÃO DA LITERATURA	19
4	FUNDAMENTAÇÃO	21
4.1	Serviço: conceito e características.....	22
4.2	Teoria das filas	27
4.2.1	Teoria das conversações telefônicas	28
4.2.2	Teoria das probabilidades.....	29
4.2.3	Sistema de filas	31
4.2.4	Distribuição de Poisson.....	33
4.2.5	Desempenho de um sistema de filas.....	36
5	ESTUDO DE CASO	38
5.1	A Empresa.....	38
5.2	Área de aplicação.....	39
5.3	Coleta de dados.....	40
5.3.1	Taxa de chegada (λ)	41
5.3.2	Velocidade do atendimento (μ).....	42
5.3.3	Tempo médio de espera em fila (TF)	45
5.3.4	Taxa de congestionamento (c)	47
6	DISCUSSÃO	50
7	CONCLUSÃO	57

1 INTRODUÇÃO

O aumento da competitividade e busca pela racionalização de recursos a fim de maximizar o lucro das empresas de serviços tem sido a porta de entrada para uma série de mudanças na gestão da produção nesse ramo. Em especial, nas empresas de tecnologia, onde o avanço tecnológico estimula o crescimento acelerado, é essencial garantir a escalabilidade dos processos produtivos, garantindo a redução e otimização dos custos, além de ampliar as vantagens competitivas.

Segundo Bondi (2000), escalabilidade é uma característica desejável em todo o sistema, em uma rede ou em um processo e que indica sua habilidade de manipular uma porção crescente de trabalho de forma uniforme ou estar preparado para crescer.

No entanto, a produção de serviços possui três características fundamentais que a diferencia das demais produções industriais, e que poderiam impedir esse processo de escalabilidade, controle produtivo e uniformização. A saber: alta diversificação da demanda, muito contato com o cliente e pouca automação.

Sabendo disso, é fundamental que cada uma das demandas recebidas passe por um processo de gestão adequado de entrada e de controle do tempo de espera do cliente para receber a finalização do serviço.

O caso analisado baseia-se em uma empresa de *Outsourcing* que está em fase de expansão de suas atividades e amadurecimento dos seus processos produtivos.

Por causa do rápido crescimento, tem enfrentado gargalos diários de suas filas de produção, gerando sobrecarga e desgastes de seus profissionais, e de seus clientes, culminando em alguns casos na perda de contratos.

1.1 Objetivo e escopo

Esse trabalho tem por objetivo avaliar a teoria das filas como ferramenta de apoio para a gestão de demandas e otimização das filas de atendimento de uma empresa de *Outsourcing*, bem como o entendimento dos principais pontos que podem gerar gargalo em sua linha de produção com a simulação dos ajustes necessários para o incremento de sua eficiência.

Possui como escopo o levantamento de dados e um ensaio para aplicação prática da teoria no setor de atendimento da empresa em questão.

2 METODOLOGIA

Segundo Severino (2000) existem vários métodos de pesquisa, tais como: pesquisa bibliográfica, pesquisa de campo, pesquisa experimental, pesquisa documental, entre outros.

Para entender o funcionamento de um sistema de filas foi realizada uma pesquisa bibliográfica e em seguida a aplicação dos conceitos obtidos em dados reais através de um estudo de caso, buscando como resultado principal a validação dessa teoria como ferramenta de apoio para a otimização da gestão das demandas e das filas de atendimento.

3 REVISÃO DA LITERATURA

Diversas literaturas a respeito do estudo e das simulações realizadas em setores com necessidades específicas foram publicadas conforme apontam as referências que serão apresentadas a seguir, mas poucas delas abordam simulações em gestão de filas na prestação de serviços de tecnologia, especialmente empresas de *Outsourcing*.

Em geral, a simulação não resolve problemas diretamente, mas evidencia a forma de trabalho de um determinado sistema de forma a demonstrar alterações em seu comportamento quando certos parâmetros são alterados. A simulação vem sendo amplamente utilizada no seguimento de serviços, para que a repetição possa gerar um padrão de execução que pode ser medido tal como é feito em uma máquina de produção.

Portanto, é possível utilizar as referências e aplicações em setores diversos como guia para o desenvolvimento de simulações para o setor de tecnologia específico e com ele simular o sistema como um todo, ainda que esse ocorra em boa parte do tempo de forma manual.

Com as literaturas de base disponibilizadas, o conceito pode ser aplicado em simulações de dados reais, que se tornam novos estudos e literaturas.

Concordando com Gonçalves (2000) que afirma que, nas empresas de prestação de serviços, o conceito de processo é de fundamental importância, uma vez que a seqüência de atividades nem sempre é visível, nem pelo cliente, nem por quem realiza essas atividades. Faz-se necessário para a simulação um desenho macro das etapas do processo envolvidos bem como seus respectivos atores para que sejam analisados todos os critérios de melhoria para aquele determinado processo.

Além disso, o próprio processo estabelecido contempla um detalhamento refinado das atividades para que de fato o ganho de produtividade seja estabelecido.

Como, via de regra, um processo de entrega de serviços passa por uma fila de atendimento é fundamental ter como base para o desenho e simulação do processo uma teoria que reflita adequadamente os conceitos de gestão de demandas e possa gerar insumos objetivos para a avaliação da produtividade de cada uma das etapas do processo.

Segundo Moraes (2001) a Teoria das filas é uma pesquisa operacional que dá um tratamento matemático ao estudo das filas, tendo como base os conceitos de probabilidade. Através desse estudo é possível determinar a capacidade e o dimensionamento adequado para a execução das atividades propostas para a entrega do serviço.

As primeiras pesquisas e trabalhos surgiram a partir da pesquisa operacional de Erlang publicada em 1909 mas se estendeu a outros ramos e áreas de aplicação a partir do término da Segunda Guerra mundial, sendo entre outros, setores de destaque na utilização dessa Teoria os de transporte, telecomunicações e transportes.

O capítulo seguinte de fundamentação associará os conceitos utilizados como premissas bem como as referências bibliográficas que respaldam o desenvolvimento deste trabalho.

4 FUNDAMENTAÇÃO

Um dos maiores desafios para o gerenciamento da demanda na empresa em questão é a existência de grande variabilidade na entrada, seja no tipo da demanda devido ao amplo portfólio de tecnologias suportadas, seja nas características específicas de cada cliente/solicitante.

A diversidade é refletida nas linhas de produção como infinitas prioridades que resultam em gargalos de outras filas de atendimento e que longo do ciclo produtivo geram filas de espera. Isso impacta diretamente no descumprimento dos acordos de nível de serviço pré-estabelecidos.

A falta de diretriz gerada pela diversidade também resulta em baixa produtividade, pois todo o esforço fica concentrado em priorizar e reorganizar a fila, quando deveria estar concentrado no atendimento como um todo.

De igual maneira, o contato contínuo com o cliente também pode resultar em gargalos operacionais, pois além do tempo gasto, o atendimento passa a ser cada vez mais “personalizado”, sendo esse tipo de atendimento inversamente proporcional à produção em escala.

Dessa forma, esse Trabalho baseia-se principalmente nas seguintes premissas:

- Na despadronização como causa da baixa produtividade, pois segundo Ganesi e Corrêa (1994), quanto maior o contato com o cliente maior a variabilidade e incerteza no serviço, resultando em menor produtividade.
- No gargalo como resultante da variabilidade, pois segundo Schemmener (1999) ambos estão diretamente ligados.
- Na necessidade de analisar a capacidade do sistema produtivo com base em indicadores objetivos para cada uma das etapas do processo.

Segundo Johnston e Clark (2002), a visão da gestão de filas envolve a identificação do serviço a ser oferecido considerando e entendendo as necessidades do cliente, a fim de assegurar que a prestação do serviço seja alinhada às expectativas do mesmo.

Apesar de ser uma consequência quase que natural da prestação de serviços, que via de regra não trabalha com estoques, a formação de filas pode gerar desconforto e descontentamento do cliente, com a possibilidade de quebra de acordos pré-estabelecidos e conseqüentemente a aplicação de penalidades que podem resultar na perda de contratos.

De acordo com os autores anteriormente citados, existem inúmeras maneiras de administrar as filas, que vão desde o aumento da capacidade produtiva até a gestão do próprio tempo de espera, fazendo com que ele pareça reduzido, diminuindo o tempo de espera.

Além disso, a configuração de agrupamento das filas de atendimento também pode contribuir, se bem administrada, para a otimização dos canais de entrega do serviço.

4.1 Serviço: conceito e características

Segundo Hargreaves, Zuanetti e Lee (2005), serviço é o resultado de pelo menos uma atividade desempenhada, necessariamente, na interface entre o fornecedor e o cliente e geralmente, é intangível.

Essa interface e contato com o cliente é a característica fundamental da entrega do serviço, porém se não for bem administrada, gera um percentual de variabilidade acima do que é possível gerenciar, acarretando perdas significativas para a organização. Isso porque uma outra característica do serviço, é que ele, por ser criado e consumido simultaneamente, não pode ser estocado, ou seja, um serviço desenhado para um determinado cliente não pode, via de regra, ser entregue a outro.

O fato de ser variável e não ser estocável evidencia uma terceira característica que é a de que o sistema de entrega de um serviço é aberto, ou seja, todo o impacto das variações de demanda é transmitido ao sistema. Por isso, é necessário um planejamento adequado da capacidade do serviço conforme a variabilidade do período, sendo que não existe um alto grau de confiabilidade do consumo real no futuro.

Apesar de todas essas características, o setor de serviços tem registrado crescimento considerado significativo face a outros setores que apresentam queda no mesmo período.

Esse crescimento comprova, através dos dados de receita apresentados na Tabela 1, que a entrega de serviços mais de que um simples bem intangível, carrega um valor agregado cada vez mais atrativo ao consumidor.

Tabela 1 – Indicadores de receita nominal do setor de serviços, segundo grupos de atividades

BRASIL - INDICADORES DE RECEITA NOMINAL DO SETOR DE SERVIÇOS, SEGUNDO GRUPOS DE ATIVIDADES
PM S - JUNHO 2013

ATIVIDADES	MÊS/IGUAL MÊS DO ANO ANTERIOR			ACUMULADO	
	TAXA DE VARIAÇÃO (%)			TAXA DE VARIAÇÃO (%)	
	ABR	MAI	JUN	NO ANO	12 MESES
BRASIL	11,6	7,6	8,6	8,4	8,9
1 - Serviços prestados às famílias	11,2	10,6	9,0	9,3	9,7
1.1 - Serviços de alojamento e alimentação	12,7	10,6	10,3	10,0	10,3
1.2 - Outros serviços prestados às famílias	3,1	10,6	1,2	4,6	6,1
2 - Serviços de informação e comunicação	10,1	5,9	7,6	6,9	6,3
2.1 - Serviços TIC	10,5	5,8	8,2	7,3	6,7
2.2 - Serviços audiovisuais, de edição e agências de notícias	7,1	6,7	3,6	3,9	4,1
3 - Serviços profissionais, administrativos e complementares	12,1	7,6	7,8	8,3	9,9
3.1 - Serviços técnico-profissionais	9,6	5,8	2,0	5,9	8,3
3.2 - Serviços administrativos e complementares	13,0	8,2	10,1	9,2	10,6
4 - Transportes, serviços auxiliares dos transportes e correio	14,0	9,9	9,8	10,8	11,3
4.1 - Transporte terrestre	16,4	10,0	11,7	11,8	11,2
4.2 - Transporte aquaviário	20,6	13,2	12,0	15,5	18,0
4.3 - Transporte aéreo	14,5	16,1	11,8	15,5	13,3
4.4 - Armazenagem, serviços auxiliares dos transportes e correio	8,5	7,5	5,5	6,9	9,7
5 - Outros serviços	7,9	2,6	11,0	5,5	6,9

Fonte: IBGE (2013)

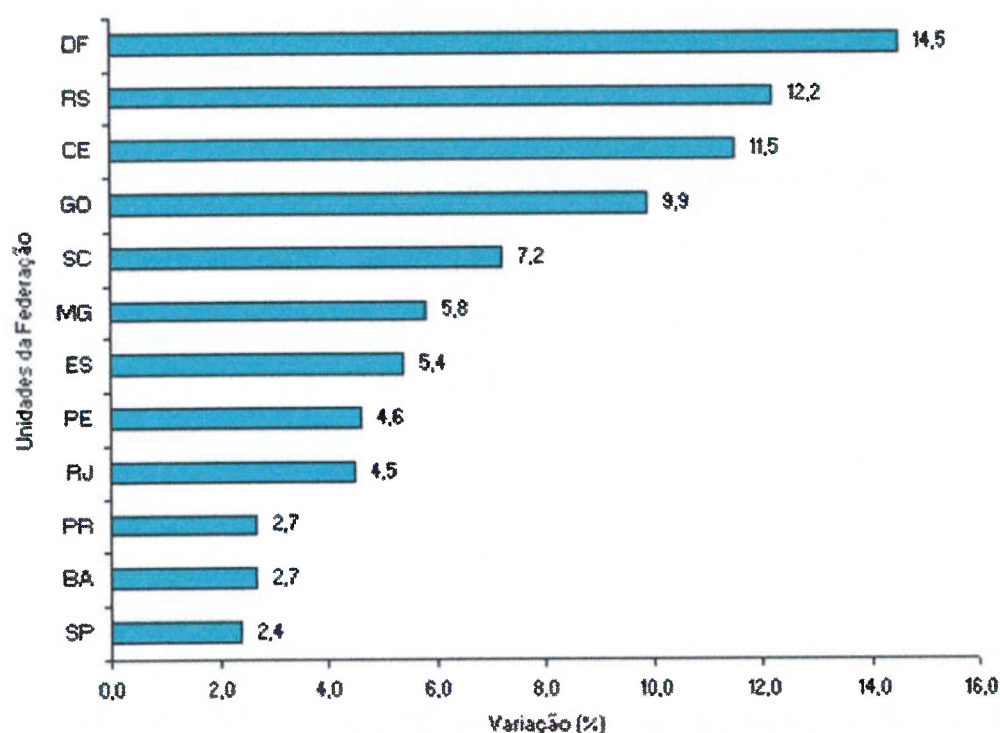
Ainda de acordo com o IBGE, especificamente o setor de serviços de informação e comunicação registra crescimento de 7,6% no mês, com destaque para os serviços

de tecnologia da informação e comunicação (TIC), com variação de 8,2%, e de 3,6% nos serviços audiovisuais, de edição e agências de notícias.

O segmento representa 31,4% da contribuição relativa no mês, contribuindo com 2,7% para a composição do índice geral, sendo que o Distrito Federal destaca-se com a maior taxa de crescimento (14,5%), seguido pelo Rio Grande do Sul (12,2%) e Ceará (11,5%).

As menores taxas foram observadas em Pernambuco e Rio de Janeiro, com média de 4,5%, Paraná e Bahia (ambas com 2,7%) e São Paulo (2,4%), conforme apresentado na figura 1.

Figura 1 – Variação mês atual/igual mês do ano anterior dos serviços de informação e comunicação, por Unidades da Federação



Fonte: IBGE (2013)

Dentro da classificação de serviços de informação e comunicação estão os serviços relacionados à tecnologia da informação que possuem as mesmas características supracitadas, bem como o crescimento constante nos últimos meses.

As tecnologias da Comunicação e da Informação estão absolutamente presentes no setor de serviços e principalmente nas organizações, que investem recursos para serem capazes de absorver as rápidas transformações e evoluções do conhecimento.

De fato, as mudanças nessa categoria têm praticamente movimentado as organizações, influenciando seu direcionamento quase como uma força econômica. Esse movimento, também chamado de revolução tecnológica, iniciou nos anos 60 após o domínio dos *mainframes*, passou pelos anos 80 com a criação dos microcomputadores e chegou aos anos 90 com a integração e circulação de informações em tempo real.

Em 1995, Rifkin apontava que as corporações começariam a estruturar o ambiente de trabalho para torná-lo compatível com a nova cultura das máquinas de alta tecnologia que estavam surgindo e Tenório, em 2007, reforçou que as estruturas organizacionais, de fato, estavam se alterando em função da revolução tecnológica.

Parte desse movimento foi impulsionado pelo desenvolvimento acelerado de *softwares* capazes de processar informações de negócio cada vez melhor e ainda, de gerar visibilidade da saúde da organização auxiliando decisivamente na gestão.

O desenvolvimento de aplicações voltadas ao negócio impulsionou uma necessidade ainda maior de suporte do desempenho e da disponibilidade dos dados nelas armazenados, e com isso o setor de infra estrutura e sustentação, dentro da tecnologia da informação passou a ser amplamente desenvolvido.

Segundo Rezende (1999) os sistemas de gestão estão intimamente ligados à essa nova cultura e certamente essa revolução impacta a organização como um todo.

Desde aquela época e mais fortemente agora, os modelos de gestão estão cada vez mais focados na redução de custos e na especialização do conhecimento do negócio. Com isso, as atividades de suporte que não são o objetivo final da empresa estão cada vez mais sendo terceirizadas.

De acordo com Marcelino (2007), terceirização é todo processo de contratação de trabalhadores por empresa interposta. Ou seja, é a relação onde o trabalho é realizado para uma empresa, mas não por ela mesma e sim contratado de maneira imediata por outra, transferindo os riscos e parte dos custos com a contratação da força de trabalho, que passam a ser denominados terceiros.

O termo terceirização, em atividades de tecnologia, também pode ser substituído por “*Outsourcing*”, uma palavra de origem inglesa, que, de acordo com Oliveira (1994) pode ser traduzida pela prestação de serviços por terceiros a alguma entidade cuja atividade principal não é aquela.

Portanto, além da transferência dos riscos de contratação que posteriormente acaba gerando redução de custos, também existe uma característica fundamental na terceirização e que vem sendo fator decisivo no processo de análise de contratação de serviços nessa modalidade, que é a especialização da contratada em atividades que não são objeto direto de ganho da empresa mas que são responsáveis pela sustentação do negócio e que indiretamente contribuem para o resultado. Por causa disso, as empresas de *Outsourcing* de Tecnologia tem ganhado cada vez mais espaço nas organizações.

De maneira geral, as empresas que oferecem esse tipo de serviço estão fundamentadas em estruturas de atendimento que funcionam com base em um escopo definido de contrato. Através desse ciclo, o negócio é sustentado e agrega valor de evolução tecnológica para a contratante.

Assim como em qualquer ciclo de entrega de produto, a entrega desse produto chamado serviço passa por uma linha de produção que se inicia com a demanda do cliente conhecido dentro desse processo como solicitante e termina com a finalização do atendimento, sendo que muitas vezes essa finalização não gera nenhuma entrega física, apenas virtual.

Por ser um ciclo de produção similar ao de uma indústria, é possível associar o processo produtivo às mesmas definições anteriormente validadas na indústria, ou

seja, os conceitos são absolutamente aplicáveis, apesar disso, as características específicas do serviço, mencionadas anteriormente, gera em muitos casos a dificuldade da escalabilidade que é, em geral, facilmente obtida em um processo industrial que tem pouca interferência do cliente.

Como dito anteriormente, a variabilidade da demanda influencia diretamente a gestão da capacidade produtiva que deve ser devidamente minimizada para que não cause impacto.

De acordo com Fitzsimmons (2004) essa variabilidade induzida pelo cliente pode ser: de capacidade, solicitação, esforço (quando espera-se que o cliente cumpra uma parte do processo) e de preferência subjetiva/pessoal em relação às de taxas de chegada.

Dependendo desse percentual de variabilidade e de organização da demanda, as solicitações podem se acumular, gerando filas de espera pela entrega do serviço.

Também de acordo com Fitzsimmons (2004), fila é uma linha de clientes que necessitam de atendimento, esperando o serviço de um ou mais prestadores.

4.2 Teoria das filas

O sistema de atendimento, bem como os problemas relacionados às filas de espera surgiu há algum tempo e pensando em soluções para esses problemas, Agner Krarkup Erlang, em 1909, iniciou o desenvolvimento de uma pesquisa que consistia na Teoria das Probabilidades e Conversações Telefônicas, provando que a teoria de Poisson era aplicável ao tráfego de chamadas telefônicas.

Em 1917, Erlang avançou em mais uma pesquisa, que possibilitou o conhecimento do cálculo de espera e de tempo perdido.

Através desse estudo, foi possível avançar não apenas na análise do fluxo do tráfego, mas também do escalonamento e da prestação de serviços, que é o foco dessa pesquisa.

De acordo com Kees (2009), a teoria das filas é usada para a análise de sistemas nos quais a ênfase recai nos indicadores de desempenho, tais como tempos de espera, tempos de conclusão e utilização de capacidade. De acordo com Davis (1999), a Teoria das Filas pode ser utilizada tanto no ambiente da manufatura quanto em serviços, sendo também uma ferramenta para o gerenciamento das operações e consequentemente uma maneira lógica de analisar os fluxos de trabalho.

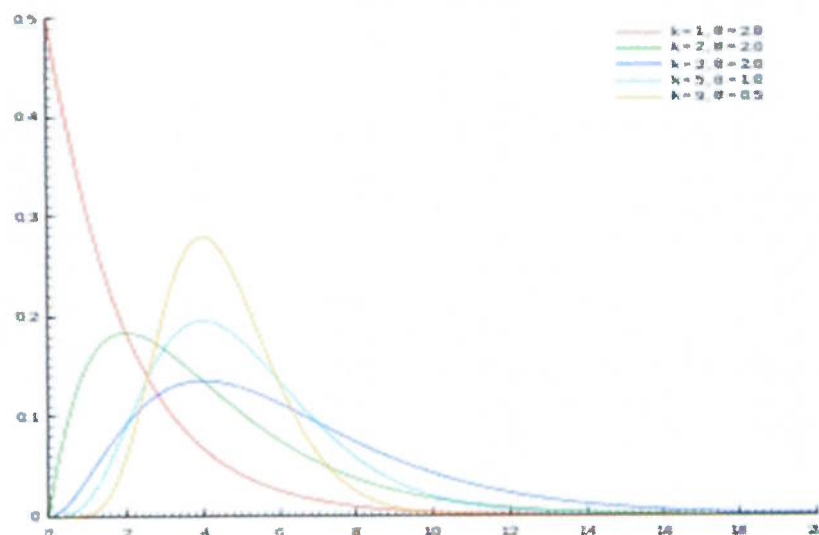
4.2.1 Teoria das conversações telefônicas

Erlang foi o primeiro a estudar os problemas relacionados à telefonia e o fez através da análise das trocas de um pequeno vilarejo, criando uma fórmula para calcular a fração das ligações que tentavam chamar alguém fora do vilarejo e que tinham que esperar pois todas as linhas estavam em uso, ou seja, calculava a taxa de congestionamento das linhas.

Sua primeira publicação *The Theory of Probabilities and Telephone Conversations*, demonstrou que demandas de ligações eram distribuídas aleatoriamente seguindo de forma aproximada a lei de distribuição de Poisson.

Essa distribuição pode ser representada através do gráfico de densidade de probabilidade conforme apresentado no gráfico 1.

Gráfico 1 – Densidade de Probabilidade



Fonte: Bussab (2011)

4.2.2 Teoria das probabilidades

Pierre Simon Laplace é um dos principais estudiosos que iniciaram o entendimento da teoria da probabilidade que permite, entre outros aspectos, calcular a ocorrência de um determinado evento em um experimento aleatório, ou seja, calcular o resultado de uma mesma ação executada várias vezes de forma aleatória.

Com base na Teoria das probabilidades e com suposições adequadas, é possível analisar todos os dados de um fenômeno e criar um modelo teórico que reproduza a distribuição de frequências tal quando esse fenômeno é observado diretamente. A reprodução desses modelos também é conhecida como modelos probabilísticos.

Esses modelos devem ser construídos adotando algumas premissas como o universo de possibilidades, de fatores favoráveis ao evento e a frequência com que esse evento pode ocorrer.

As probabilidades podem ser condicionais e subjetivas.

- A probabilidade condicional refere-se à probabilidade de um evento ocorrer sabendo que outro evento já ocorreu, ou seja, a probabilidade de ocorrer um evento depende da ocorrência de outro evento.

A probabilidade de A condicionada por B (ou dado B , ou sabendo que B) é definida, sabendo que $P(B) > 0$, pela fórmula:

$$P(A|B) = P(A \cap B) / P(B)$$

Assim, a probabilidade de A muda após o evento B ter acontecido. Isso porque o resultado de A é uma das possibilidades de B .

- A probabilidade subjetiva leva em consideração que cada indivíduo, baseado em informações anteriores e nas opções particulares a respeito do evento pode ter uma resposta para o evento, ou seja, o fato de ter ocorrido um determinado evento de uma determinada maneira pode influenciar diretamente a probabilidade de ocorrência de um novo evento se esse tiver interferência do indivíduo.

Independentemente do tipo de probabilidade, sempre existiram variáveis que farão a composição do resultado, essas podem ser Quantitativas ou Qualitativas

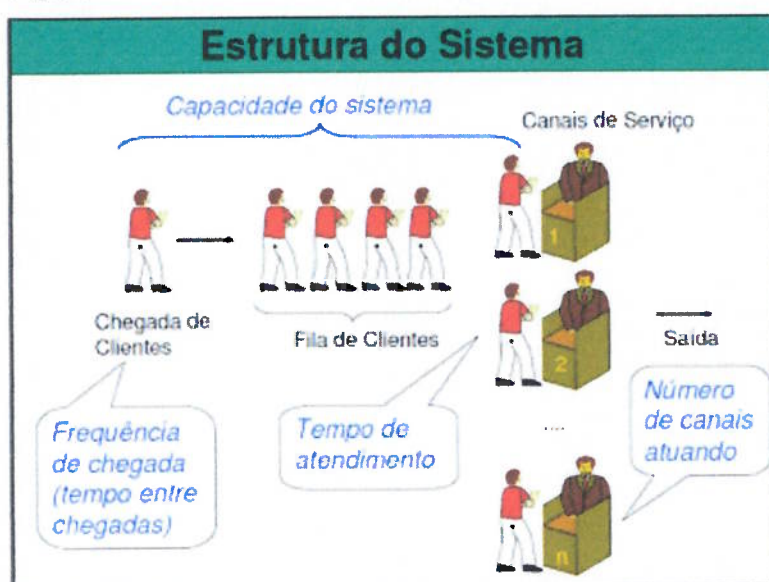
- Variáveis quantitativas: apresentam valores numéricos que fazem sentido, sendo esses discretos (assumem valores inteiros) ou contínuos (assumem valores fracionados).
- Variáveis qualitativas: não representam valores numéricos, mas sim de categorias e/ou classificação de indivíduos, sendo nominais (sem ordenação entre as categorias) ou ordinais.

4.2.3 Sistema de filas

A Teoria das filas baseia-se na existência de um sistema de filas, ou seja, um cliente chega, espera pelo serviço e sai do sistema após ser atendido conforme ilustrado na figura 2.

Um sistema de filas ou simplesmente uma fila, somente é formado quando a procura por um serviço é maior do que capacidade de seu provimento.

Figura 2 – Ilustração de um sistema de filas



Fonte: Instituto Paraclete (2013)

Para o segmento de serviços, a existência de uma fila em si não pode ser imediatamente indicada como um problema a ser resolvido sem que algumas variáveis sejam analisadas. Essas vão desde o custo da produção do serviço até a expectativa do cliente a respeito do tempo do recebimento do serviço, chamado de grau de paciência do cliente.

De maneira geral, apenas se essas perspectivas não estiverem devidamente equilibradas é que a fila começa a ser um problema para a manutenção do serviço.

Portanto, o equilíbrio das perspectivas do fornecedor e do cliente deve ser realizado através do mapeamento completo do processo de entrega do serviço, desde a solicitação até a entrega, e com ele todos os pontos de interface.

A otimização de um sistema de filas pode ser feito utilizando a Teoria das Filas desde que o sistema esteja devidamente definido conforme as seguintes perspectivas:

- Processo de chegada
- População de usuários
- Número de servidores
- Disciplina de atendimento
- Distribuição do tempo de serviço
- Capacidade do sistema

4.2.3.1 Processo de chegada

Através da análise do processo de chegada, é possível identificar como e com que frequência os serviços são solicitados.

Exceto os sistemas automatizados que podem garantir uma taxa constante de entradas no sistema, nos demais, geralmente, existe algum grau de variabilidade e por causa disso o cálculo de chegada deve ser feito com base em uma distribuição de probabilidade. Para que esse comportamento seja conhecido, é importante identificar qual distribuição descreve adequadamente os tempos de intervalo entre as chegadas das demandas.

De acordo com Bussab (2011), existem muitas distribuições de probabilidade que podem ser utilizadas para a análise da entrada em um sistema de filas, tais como a de Erlang, a Hipere exponencial, a Arbitrária e a de Poisson, sendo essa última a mais comumente utilizada, pois em geral as empresas possuem mais de um cliente com demandas variáveis, logo a taxa de chegada acompanha a demanda desses clientes e, portanto, também é variável conforme apresentado no tópico 3.2.1.

Quanto maior a variabilidade, mais difícil torna-se prever a capacidade adequada do sistema e, portanto, a formação de filas.

4.2.4 Distribuição de Poisson

A distribuição de Poisson foi criada por Siméon-Denis Poisson e publicada, junto com a sua teoria a respeito da probabilidade em 1838. A publicação teve como principal foco certas variáveis aleatórias que contavam, entre outras coisas, o número de ocorrências contínuas que ocorriam durante um intervalo de tempo determinado.

A probabilidade de que existam exatamente o número k (inteiro não negativo) de ocorrências pode ser obtido através da fórmula:

$$f(k; \lambda) = (e^{-\lambda} \lambda^k) / k!$$

- e é base do logaritmo natural ($e = 2.71828...$)
- $k!$ é o fatorial de k ,
- λ é um número real, igual ao número esperado de eventos que ocorrem num dado intervalo de tempo

Essa distribuição pode ser observada em diversas situações, dentre elas as chegadas de clientes em uma fila de atendimento que ocorrem através de um processo estocástico, ou seja, sem a dependência de um evento anterior, de forma aleatória em um tempo determinado de chegada.

4.2.4.1 População de usuários

Além da taxa de chegada, a definição das características da população também é fundamental para que o sistema de filas seja devidamente analisado. Essa população entrante pode ser definida como finita ou infinita.

- População finita, como o próprio nome diz, é uma população que possui um número limitado de tamanho, ou seja, existe um número limitado de clientes que podem ingressar no sistema. Esse tipo de população define diretamente a

probabilidade de ocorrência da solicitação de um serviço, reduzindo o indicador cada vez que a população diminui por qualquer motivo que seja, e aumentando quando o a população retorna ao tamanho normal.

- População infinita é a que possui um tamanho tão significativo em relação ao sistema, que o seu tamanho não influencia diretamente as probabilidades de chegada no sistema, ou seja o ritmo de chegada não é afetado pelo número de pessoas do sistema.

4.2.4.2 Número de servidores

O número de servidores também pode ser identificado como o número de canais de serviço disponíveis para atender ao cliente de forma paralela.

Quando existe mais de um canal disponível para atendimento, existem duas formas de agrupamento no sistema, sendo o de fila única ou de múltiplas filas, ou seja, uma fila para cada canal de atendimento.

4.2.4.3 Disciplina de atendimento

A disciplina de atendimento define como os clientes sairão da fila de espera para serem atendidos.

Dentre as disciplinas existentes estão:

- FCFS (*First Come, First Served*): primeiro a chegar, primeiro a ser atendido.
- FIFO (*First In, First Out*): primeiro a entrar, primeiro a sair.
- LCFS (*Last Come, First Served*): último a chegar, primeiro a ser atendido
- LIFO (*Last In, First Out*): último a chegar, primeiro a sair.
- Fila com prioridade: cada cliente possui uma prioridade e podem ter preferencia no atendimento.

- *Round-robin*: cada cliente recebe uma fatia de tempo de canal de atendimento. Se ao final do tempo destinado o serviço não tiver sido concluído o cliente deixa de ser atendido e o próximo começa a ser atendido. Posteriormente o serviço interrompido é retomado, em ciclos definidos de tempo e de forma circular.

4.2.4.4 Distribuição do tempo de serviço

Da mesma forma que é importante analisar o processo de entrada, também é importante identificar a distribuição do tempo de serviço, considerando que os serviços podem ser simples ou batch.

O serviço é considerado *batch* quando a entrada no sistema é feita através de um lote de serviços ou de tarefas enfileiradas, sendo que o sistema só processa o próximo atendimento após o término completo do lote anterior.

A distribuição do tempo mede essencialmente a capacidade em tempo que um servidor ou canal de serviço leva para executar essa atividade e, portanto, estará livre para iniciar a próxima

Essa distribuição está associada diretamente à capacidade individual do servidor/analista, já que fatores como sobrecarga e conhecimento apropriado são variáveis altamente relevantes.

4.2.4.5 Capacidade do sistema

A capacidade está relacionada com o volume máximo de clientes que o sistema suporta, incluindo os que estão sendo atendidos e os que estão em fila(s) de espera.

- Sistema finito: quando um sistema suporta um número limitado de clientes ele, ou seja, quando sua capacidade máxima é atingida o cliente é rejeitado e não consegue ser inserido na espera.
- Sistema infinito: quando não existe a limitação de suporte ao número de clientes.

4.2.4.6 Notação

A análise das características do sistema permite adotar a notação idealizada por Kendall (1951) que define de forma clara e objetiva as variáveis que compõem o sistema, sendo dispostas como $A/S/m/K/N/Q$, onde:

- A: descreve o processo de chegada
- S: descreve a distribuição dos tempos de serviço
- m: representa o número de servidores ou atendentes
- K: representa a capacidade do sistema
- N: define o tamanho da população ou clientes do sistema
- Q: define a disciplina da atendimento.

4.2.5 Desempenho de um sistema de filas

Para que seja possível o correto entendimento do funcionamento de um sistema produtivo e em seguida estabelecer propostas de melhoria para esse sistema é necessário basear-se em indicadores que tenham variáveis bem definidas e critérios objetivos.

No cálculo de desempenho de filas são utilizadas as seguintes variáveis:

- c: número de atendentes do sistema
- λ : taxa média de chegada de clientes

- μ : taxa média de atendimento por atendente
- ρ : taxa de utilização do sistema
- Wq : tempo do cliente em fila
- Ws : tempo do cliente durante o atendimento
- W : tempo total de um cliente no sistema
- Lq : número de clientes na fila ($Lq = \lambda * Wq$)
- Ls : número de clientes em atendimento ($Ls = \lambda * Ws$)
- L : número total de clientes ($L = \lambda * W$)

O cálculo das variáveis, em alguns casos, depende do modelo de sistema, ou seja, se existe um ou mais servidores ou analistas disponíveis, se as saídas ocorrem conforme o fluxo de entrada ou mesmo da capacidade produtiva do servidor ou analista.

Conforme mencionado anteriormente, o sistema pode ser modelado através da notação de Kendall.

Dessa forma, após a formatação do modelo de sistema, o alinhamento dos conceitos envolvidos e o conhecimento das variáveis foi possível aplicar a Teoria das filas como estudo de caso, sendo apresentados os resultados nos tópicos seguintes.

5 ESTUDO DE CASO

5.1 A Empresa

Trata-se de uma empresa Tecnologia da Informação fundada em 1999 que é referência na prestação de serviços de *Outsourcing* para ambientes de missão crítica.

Entende-se por ambientes de missão crítica, aqueles responsáveis pela parte central da organização e que se indisponíveis, geram principalmente a perda financeira e de lucratividade ou impedem o funcionamento pleno da organização.

Essa organização tem em seu portfólio mais de 250 clientes de médio e grande porte de diversos segmentos, e apesar de possuir parcerias com fabricantes e integradores, é considerada isenta de vínculo. Possui ainda conhecimento em diversas plataformas de tecnologia, portanto, considerada multiplataforma.

Foi responsável pelo desenvolvimento de um *software* amplamente utilizado pelo mercado em transações de cartões de crédito e também pela implementação de governança de TI em vários de seus clientes.

Apesar de ser uma empresa voltada para o mercado nacional, participou de projetos internacionais o que ajudou a solidificar seu nome também em outros mercados.

Foi fundada por especialistas de TI e contou com o investimento de empresas multinacionais no início da sua fundação e no ano passado seu controle foi adquirido pela maior empresa de *Data Center* brasileira, objetivando com isso, ser o maior grupo de tecnologia do país.

Com a aquisição do controle, outras empresas do grupo com ramos de interesse convergentes foram unificadas para formar uma única unidade, que oferta serviços de *Outsourcing* e soluções.

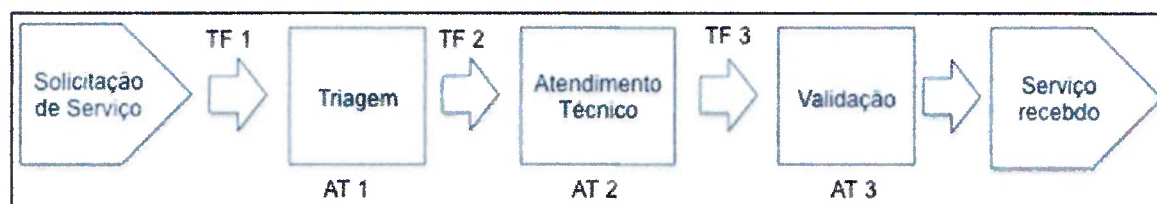
O horizonte da unidade é o de crescimento e ganho de mercado. Para que isso seja possível, foram incluídas metas de crescimento e geração de receita bastante agressivas em comparação com o setor.

Apesar da união com demais empresas do grupo, essa unidade preserva a visão da fundação, que incluiu ser referência em excelência técnica, garantir a satisfação dos clientes, gerar valor para o acionista de maneira ética e com responsabilidade social e, principalmente, promover o desenvolvimento contínuo dos seus funcionários.

5.2 Área de aplicação

O produto entregue pela unidade em questão é essencialmente o serviço de suporte técnico e para que a entrega seja possível é necessário que esse serviço passe pelo fluxo de atendimento que segue desde o registro da demanda pelo cliente até a conclusão e validação do atendimento, como mostra na figura 3.

Figura 3 – Fluxo de entrega de serviço



Onde:

TF = Tempo em fila

AT = Atendimento/Tempo de atendimento

Para que a demanda passe pelo fluxo, é necessário que seja identificada a fila correta de atendimento através do mapeamento dos requisitos e necessidades informadas pelo cliente. Essa etapa é conhecida como etapa de triagem.

Com essa etapa finalizada, essa demanda é encaminhada para a fila de atendimento técnico.

Sabendo que as linhas de atendimento são diversas e que concorrem entre si, faz-se necessária a avaliação de probabilidade de utilização dessas linhas, para que o dimensionamento de recursos seja feito corretamente evitando gargalo nas filas e o impacto na satisfação do cliente.

Após o atendimento técnico, a solicitação passa pela etapa de validação do atendimento, e se a solicitação estiver completamente atendida, é encerrada e o serviço considerado entregue.

A área de aplicação desse trabalho é o processo de entrega de uma requisição de serviço considerando os tempos de atendimento das equipes envolvidas e as etapas de espera para que seja possível o dimensionamento adequado das equipes.

A empresa em questão realiza, aproximadamente, 15.000 atendimentos por mês dentro das várias linhas de serviço existentes, porém, para este trabalho, será analisado apenas a entrega do serviço de ativação de conta de correio, que corresponde a uma amostra da produção inteira da unidade.

Os conceitos de sistema de filas estarão aplicados através da análise de desempenho do sistema.

5.3 Coleta de dados

Os dados obtidos para esse trabalho partiram do sistema utilizado pela empresa em questão, que contabiliza os atendimentos realizados e mantém todas as informações de tempos, atendentes, solicitantes e tipos de solicitações. A base utilizada foi do ano de 2013.

As informações foram disponibilizadas através de um arquivo em planilha que possibilitou as simulações que serão demonstradas nos próximos tópicos.

Como no processo analisado existem três etapas de atendimento (triagem, atendimento técnico e validação), o sistema será dividido em três subsistemas para que seja possível avaliar a espera no sistema como um todo.

5.3.1 Taxa de chegada (λ)

De acordo com o apontamento do sistema de atendimento, foi possível identificar a taxa de chegada dos clientes na fila.

Os dados obtidos foram:

- Número de solicitações = 2686
- Intervalo de tempo = 160 horas (8 horas por dia durante 20 dias)

Sabendo que a taxa de chegada é calculada segundo a fórmula:

$$\lambda = \text{número de solicitações} / \text{intervalo de tempo}$$

$$\lambda = 2686 / 160$$

$$\lambda = 17$$

a taxa de chegada no período analisado foi de, aproximadamente, 17 solicitações por hora.

A solicitação deve, necessariamente, passar pelos três subsistemas, portanto, a mesma taxa foi aplicada para os três.

5.3.2 Velocidade do atendimento (μ)

Para o cálculo da velocidade de atendimento foi necessário segregar as etapas em três, conforme justificado anteriormente, pois os tempos de atendimento em cada uma delas são significativamente diferentes entre si.

5.3.2.1 Velocidade de atendimento na etapa Triagem

Os dados obtidos para os cálculos dessa etapa foram:

- Número de atendentes para a etapa de triagem = 2
- Tempo de execução da atividade = 0,25 horas
- Intervalo de tempo = 160 horas (8 horas por dia durante 20 dias)

Sabendo que a velocidade de atendimento (μ) é calculada pela fórmula:

$$\mu = (\text{tempo de execução da atividade} / \text{intervalo de tempo}) * (\text{quantidade de analistas})$$

$$\mu = (160/0,25)*2$$

$$\mu = 1280$$

a capacidade de atendimento por mês para a etapa triagem, portanto, é de 1280 solicitações, pois o intervalo de tempo (horas disponíveis de trabalho) dividido pelo tempo de atividade, multiplicado pela quantidade de analistas é igual a 1280.

Consolidando os dados obtidos anteriormente e os atuais:

- Número de atendentes = 2
- Tempo de execução da atividade = 0,25 horas
- Intervalo de tempo = 160 horas (8 horas por dia durante 20 dias)
- Velocidade de atendimento por mês = 1280

5.3.2.2 Velocidade de atendimento na etapa atendimento técnico

Os dados obtidos para os cálculos dessa etapa foram:

- Número de atendentes para a etapa atendimento técnico = 4
- Tempo de execução da atividade = 0,5 horas
- Intervalo de tempo = 160 horas (8 horas por dia durante 20 dias)

Sabendo que a velocidade de atendimento (μ) é calculada pela fórmula:

$$\mu = (\text{tempo de execução da atividade} / \text{intervalo de tempo}) * (\text{quantidade de analistas})$$

$$\mu = (160/0,5)*4$$

$$\mu = 1280$$

Com isso, é possível concluir que a velocidade na etapa de atendimento técnico no mês é de 1280.

Consolidando os dados obtidos anteriormente e os atuais:

- Número de atendentes = 4
- Tempo de execução da atividade = 0,5 horas
- Intervalo de tempo = 160 horas (8 horas por dia durante 20 dias)
- Velocidade de atendimento por mês = 1280

5.3.2.3 Velocidade de atendimento na etapa Validação

Os dados obtidos para os cálculos dessa etapa foram:

- Número de atendentes para a etapa validação = 3
- Tempo de execução da atividade = 0,3 horas
- Intervalo de tempo = 160 horas (8 horas por dia durante 20 dias)

Sabendo que a velocidade de atendimento (μ) é calculada pela fórmula:

$$\mu = (\text{tempo de execução da atividade} / \text{intervalo de tempo}) * (\text{quantidade de analistas})$$

$$\mu = (160/0,3)*3$$

$$\mu = 1600$$

Com isso, é possível concluir que a velocidade na etapa de validação no mês é de 1.600.

Consolidando os dados obtidos anteriormente e os atuais:

- Número de atendentes = 3
- Tempo de execução da atividade = 0,3 horas
- Intervalo de tempo = 160 horas (8 horas por dia durante 20 dias)
- Velocidade de atendimento por mês = 1600 (Número de analistas x Número de horas / tempo de atividade)

Após a análise individual das taxas de chegada e da capacidade produtiva (Velocidade de atendimento) das equipes, os resultados foram novamente analisados com base no quadro resumo disposto a seguir.

Quadro 1 – Quadro resumo de análise Taxa de chegada e Velocidade de atendimento

Situação atual	A	B	C	D	Número de atendimentos realizados por hora total (= D x A)	Número de atendimentos realizados por mês (D x B x A)
	Número analistas	Horas disponíveis de trabalho no mês	Tempo gasto ao executar 1 vez a atividade (em horas)	Número de atendimentos realizados por hora por analista (1/C)		
Triagem	2	160	0,25	4	8	1280
Atendimento	4	160	0,5	2	8	1280
Validação	3	160	0,3	3	10	1600

De acordo com o que foi apresentado anteriormente o volume médio real de solicitações no mesmo período é de 2686, sendo a taxa de chegada (λ) de 17 solicitações por hora.

Com base no número de atendimentos realizados por hora, é possível identificar uma diferença importante entre a demanda de solicitações e a capacidade produtiva em

todas as etapas. Como já dito, um dos principais motivos da formação de filas de espera é a capacidade produtiva ser menor que a demanda pelo atendimento.

5.3.3 Tempo médio de espera em fila (TF)

Como já foi caracterizado que existe, de fato, um sistema de filas formado com os dados obtidos, foi possível calcular o tempo médio em que o cliente permanece nas filas e, portanto, no sistema.

5.3.3.1 Tempo médio de espera em fila na etapa triagem (TF)

Os dados obtidos para os cálculos dessa etapa foram:

- Tempo de espera na etapa de triagem(λ): 17
- Velocidade de atendimento (μ): 8

Sabendo que a velocidade de atendimento (μ) é calculada pela fórmula:

$$\begin{aligned}TF &= (\lambda / \mu)(\mu - \lambda) \\TF &= (17/8)(8 - 17) \\TF &= 16\end{aligned}$$

Com isso, é possível concluir que o tempo médio de fila na etapa de triagem é de 16 horas

5.3.3.2 Tempo médio de espera em fila para a etapa atendimento (TF)

Os dados obtidos para os cálculos dessa etapa foram:

- Tempo de espera na etapa de triagem(λ): 17
- Velocidade de atendimento (μ): 8

Sabendo que a velocidade de atendimento (μ) é calculada pela fórmula:

$$\begin{aligned} TF &= (\lambda / \mu)(\mu - \lambda) \\ TF &= (17/8)(8-17) \\ TF &= 16 \end{aligned}$$

Com isso, é possível concluir que o tempo médio de fila na etapa de atendimento é de 16 horas.

5.3.3.3 Tempo médio de espera em fila para a etapa validação (TF)

Os dados obtidos para os cálculos dessa etapa foram:

- Tempo de espera na etapa de triagem(λ): 17
- Velocidade de atendimento (μ): 10

Sabendo que a velocidade de atendimento (μ) é calculada pela fórmula:

$$\begin{aligned} \text{TF} &= (\lambda / \mu)(\mu - \lambda) \\ \text{TF} &= (17/10)(10-17) \\ \text{TF} &= 16 \end{aligned}$$

Com isso, é possível concluir que o tempo médio de fila na etapa de validação é de 16 horas.

Ou seja, somando as três etapas, o cliente espera, aproximadamente, 50 horas para ser atendido, conforme detalhado no quadro 2.

Quadro 2 - Quadro resumo de análise do tempo médio em fila

Situação atual	A	B	C	D	Número de atendimentos realizados por hora total (= D x A)	Número de atendimentos realizados por mês(Dx B x A)	Tempo médio em fila
	Número analistas	Horas disponíveis de trabalho no mês	Tempo gasto ao executar 1 vez a atividade (em horas)	Número de atendimentos realizados por hora por analista (1/C)			
Triagem	2	160	0,25	4	8	1280	-16,7
Atendimento	4	160	0,5	2	8	1280	-16,7
Validação	3	160	0,3	3	10	1600	-16,8
TOTAL							-50,3

5.3.4 Taxa de congestionamento (c)

A análise dos valores dos tempos em fila mostraram que existe um congestionamento das filas de atendimento analisadas, ou seja, a capacidade de produção não está sendo suficiente para suprir a demanda. Com isso cabe a análise da taxa de congestionamento que servirá de base para as propostas de solução de redução do tempo de espera.

5.3.4.1 Taxa de congestionamento para a etapa triagem

Os dados obtidos para os cálculos dessa etapa foram:

- Taxa de chegada (λ): 17
- Taxa de equilíbrio: 1
- Velocidade de atendimento (μ): 8

Sabendo que a taxa de congestionamento (c) é calculada pela fórmula:

$$\begin{aligned}c &= \text{Taxa de chegada} / (\text{Taxa de equilíbrio} \times \text{Velocidade de atendimento}) \\c &= 17 / (1 \times 8) \\c &= 2,1\end{aligned}$$

Com isso, é possível concluir que a taxa de congestionamento na etapa de triagem é de 2,1 horas.

5.3.4.2 Taxa de congestionamento para a etapa atendimento técnico

Os dados obtidos para os cálculos dessa etapa foram:

- Taxa de chegada (λ): 17
- Taxa de equilíbrio: 1
- Velocidade de atendimento (μ): 8

Sabendo que a taxa de congestionamento (c) é calculada pela fórmula:

$$c = \text{Taxa de chegada} / (\text{Taxa de equilíbrio} \times \text{Velocidade de atendimento})$$

$$c = 17 / (1 \times 8)$$

$$c = 2,1$$

Com isso, é possível concluir que a taxa de congestionamento na etapa de atendimento técnico é de 2,1 horas.

5.3.4.3 Taxa de congestionamento para a etapa Validação

Os dados obtidos para os cálculos dessa etapa foram:

- Taxa de chegada (λ): 17
- Taxa de equilíbrio: 1
- Velocidade de atendimento (μ): 10

Sabendo que a taxa de congestionamento (c) é calculada pela fórmula:

$$c = \text{Taxa de chegada} / (\text{Taxa de equilíbrio} \times \text{Velocidade de atendimento})$$

$$c = 17 / (1 \times 10)$$

$$c = 1,7$$

Com isso, é possível concluir que a taxa de congestionamento na etapa de validação é de 1,7 horas.

Ou seja, os valores obtidos para a taxa de congestionamento para as etapas de Triagem, Atendimento técnico e Validação foram, respectivamente: 2,1, 2,1 e 1,7 conforme demonstrado na quadro 3.

6 DISCUSSÃO

Considerando os dados obtidos na pesquisa, é possível concluir que um dos principais motivos para o gargalo atual no atendimento está relacionado ao número de analistas/atendentes disponíveis para a execução das atividades.

Com a disponibilidade e capacidade atual, um atendimento que deveria demorar apenas 1,08 horas é atendido em 50 horas, sendo mais de 95% do tempo da solicitação no sistema dedicado ao tempo de espera em fila.

Especificamente para o serviço que foi analisado, o solicitante tem a expectativa de espera (Grau de Paciência) de até 8 horas considerando o ciclo completo (todas as fases do atendimento), portanto, o tempo de espera atual com os recursos disponíveis ultrapassa em, aproximadamente, 42 horas a expectativa do cliente.

Como qualquer serviço contratado, parte do valor depositado para a sua aquisição está relacionado ao tempo em que o serviço será entregue. Em alguns casos, costuma-se pagar um valor superior simplesmente pela garantia de entrega do serviço mais rápido do que o padrão. Independentemente do acréscimo pela urgência, o cliente que contrata um serviço tem como principal expectativa a entrega dentro do prazo contratado além da qualidade.

Analisando a taxa de congestionamento é possível observar que para cada uma das etapas faltam recursos e/ou as atividades não estão sendo realizadas no tempo ideal, ou seja, precisam ser otimizadas.

Como a empresa analisada estabeleceu indicadores de produtividade que comprovam que o tempo executado pelos analistas está adequado para a atividade, será analisado apenas a inclusão ou não de analistas/servidores no sistema.

Para que isso seja possível e eficiente, ao incluir um recursos é necessário analisar novamente todas as variáveis para que o sistema não fique ocioso e não gere aumento de custos desnecessários para a produção.

A ação, portanto, de incluir recursos como solução deve ser analisada sob dois aspectos: aumento do custo pela entrega do serviço e ociosidade ainda que temporária.

Para uma empresa de serviços em que a matéria prima e os custos de produção são essencialmente pessoas, qualquer ociosidade pode impactar em queda de margem/lucratividade e, em alguns casos, colocar em risco a saúde financeira da empresa. Por outro lado, o fato de não atender os clientes dentro do tempo contratado pode gerar perda de contrato, conseqüentemente perda de lucratividade e risco financeiro.

Sendo assim, apesar do problema ter sido facilmente identificado através dos cálculos realizados, a aplicação da solução deve passar por uma análise detalhada das variáveis antes da decisão pela aplicação.

De acordo com COGAN (1998) e demais literaturas relacionadas à fila de espera, existem também sugestões alternativas para minimizar o tempo de espera e com isso deixar os clientes menos insatisfeitos, porém como no caso abordado trata de uma solicitação que fica em um sistema de atendimento aguardando atendimento, e não um cliente aguardando pessoalmente cabe a busca por outras soluções que reflitam diretamente na redução do tempo em fila.

Conforme mencionado anteriormente, para o serviço analisado, o cliente tem a expectativa de espera de até 8 horas e esse é o custo pelo qual ele está disposto a pagar, ou seja, nesse caso o cliente não está disposto a pagar pelo custo de não ter nenhuma fila, mas sim de ter uma fila que proporcione a entrega do que foi solicitado em até 8 horas.

Sendo assim, a simulação da taxa de congestionamento levará em consideração um tempo médio de espera de 8 horas no sistema como um todo.

Ao iniciar a simulação dessa redução, além da premissa de expectativa de espera, também foi levado em consideração o custo por analista, ou seja, o valor gasto para o atendimento de uma solicitação.

Apesar da taxa de congestionamento apontar para a inclusão de analistas em todas as fases para que o sistema esteja equilibrado, o custo desses analistas não é equivalente.

Se a principal restrição para a solução é justamente o custo, avaliar qual tipo de analista será incluído no sistema certamente fará toda a diferença no custo total do serviço.

De acordo com os dados enviados pela empresa analisada, os analistas de triagem e validação tem o custo 40% menor que os analistas técnicos, por isso foi avaliado a inclusão e concentração de analistas de triagem e validação em detrimento dos analistas técnicos.

Para o cliente, as divisões internas de etapas é desconhecida, ou seja, é uma variável da organização que não tem influência alguma na expectativa de entrega do cliente. Se o tempo total for reduzido, ainda que em alguma etapa o resultado do tempo em fila esteja acima do ideal, o reflexo para o cliente será absolutamente positivo se igual ou acima do Grau de Paciência do cliente.

Diante disso, foram simuladas algumas situações de inclusão de analistas em cada uma das etapas, mantendo a premissa que a etapa de atendimento é a que tem o custo maior devido ao custo dos analistas.

Após as simulações, não foi possível chegar a um equilíbrio completo em que o grau de paciência seria atendido e o sistema permaneceria em todo o tempo produzindo.

A formação que mais se aproximou do tempo total de 8 horas foi a que simulou a inclusão de 4 analistas na etapa triagem, 1 analista na etapa atendimento e 3 analistas na etapa validação.

Nessa formação, o número total de analistas nas etapas de triagem, atendimento e validação foram respectivamente: 6, 5, 6. Essa formação resultou em um tempo e atendimento total de 6,9 horas, 1,1 horas mais rápido do que a expectativa do cliente.

Analisando superficialmente apenas sob esse aspecto, seria possível afirmar que esse seria o tempo de ociosidade do sistema, porém é bastante importante considerar as variáveis de perda de produtividade e até mesmo de picos de demanda para obter o valor correto.

Sendo assim, se for considerado, 20% de perda durante o processo, por exemplo, o tempo disponível seria suficiente para manter o sistema em equilíbrio sem afetar a percepção de entrega do cliente, da mesma forma que ser fosse necessário aumentar a capacidade produtiva em 6% o tempo disponível também seria suficiente, ou seja, a formação proposta cobre a demanda atual e um crescimento futuro.

Os dados obtidos na simulação final podem ser analisados através do quadro 4.

Quadro 4 – Simulação final de equilíbrio do sistema

Situação atual	A	B	C	D	E	F	G
	Número analistas	Horas disponíveis de trabalho no mês	Tempo gasto ao executar 1 vez a atividade (em horas)	Número de atendimentos realizados por hora por analista (1/C)	Número de atendimentos realizados por hora total (= D x A)	Número de atendimentos realizados por mês (D x B x A)	Tempo médio em fila
Triagem	6	160	0,25	4	24	3840	5,0
Atendimento	5	160	0,5	2	10	1600	-11,9
Validação	6	160	0,3	3	20	3200	1,1
TOTAL							-6,95

Durante a análise dos dados obtidos, foi identificada uma nova variável que de igual maneira deveria ser garantida.

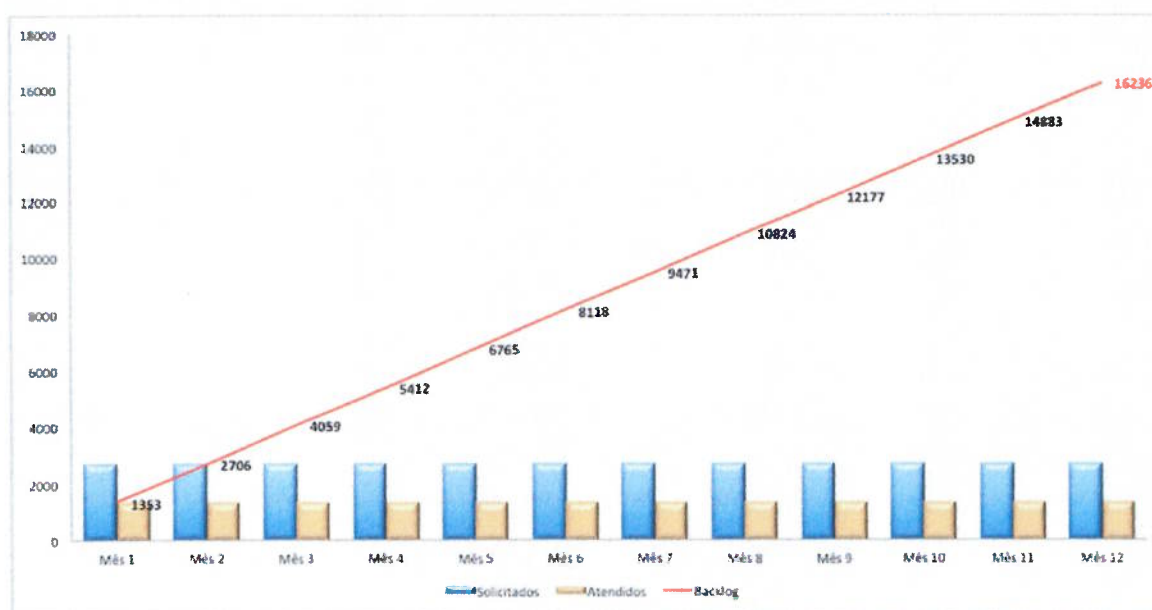
Essa variável, chamada de *Backlog*, é responsável por medir o volume de atendimentos acumulados em fila em um determinado período de tempo. Esse acúmulo, se não for bem administrado, pode gerar aumento no tempo total de atendimento, consequentemente, o desequilíbrio do sistema.

Na situação atual, a empresa analisada tem a capacidade de atendimento de 1.333 chamados por mês, esse número foi obtido através da multiplicação das horas disponíveis de trabalho dos 9 analistas dividido pelo tempo de atividade gasto no total.

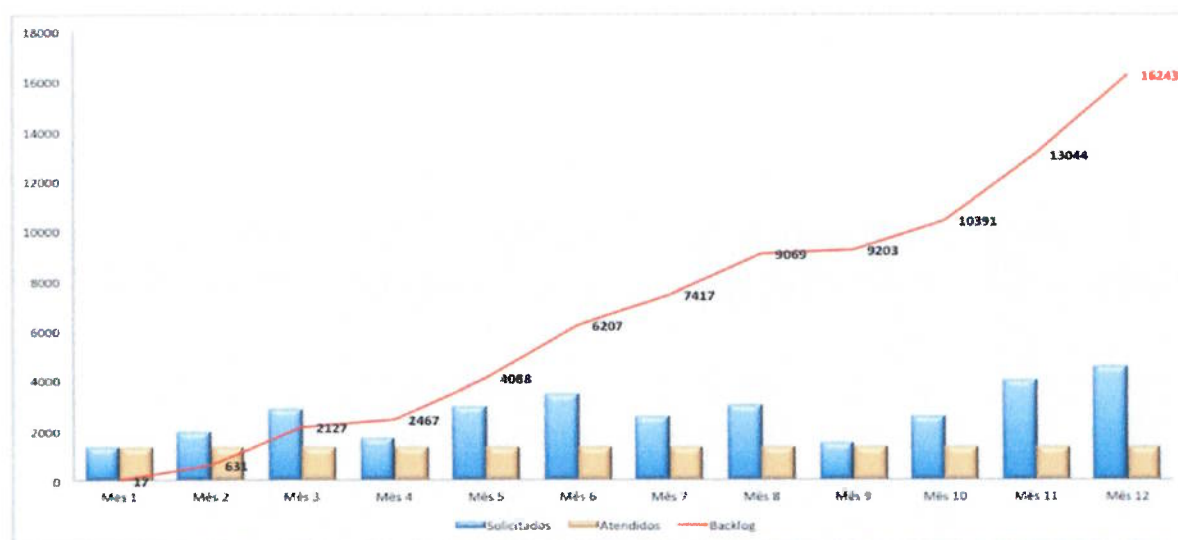
Essa capacidade, conforme já demonstrado não é suficiente para cobrir a demanda, já que o volume de solicitações é de 2.686 em média.

O sistema analisado é aberto, ou seja, as solicitações não são impedidas de entrarem, portanto, o volume de solicitações não atendidas permanece em fila.

Considerando uma demanda constante, assim como a capacidade produtiva, o cálculo do acúmulo mensal das solicitações não atendidas ao longo de 12 meses, resulta em um volume de *backlog* de 16.236 ao final de 12 meses, ou seja, um número 12 vezes maior do que a capacidade produtiva, conforme pode ser visualizado no gráfico 2.

Gráfico 2 – Evolução do *Backlog* em 12 meses. Situação atual. Demanda constante

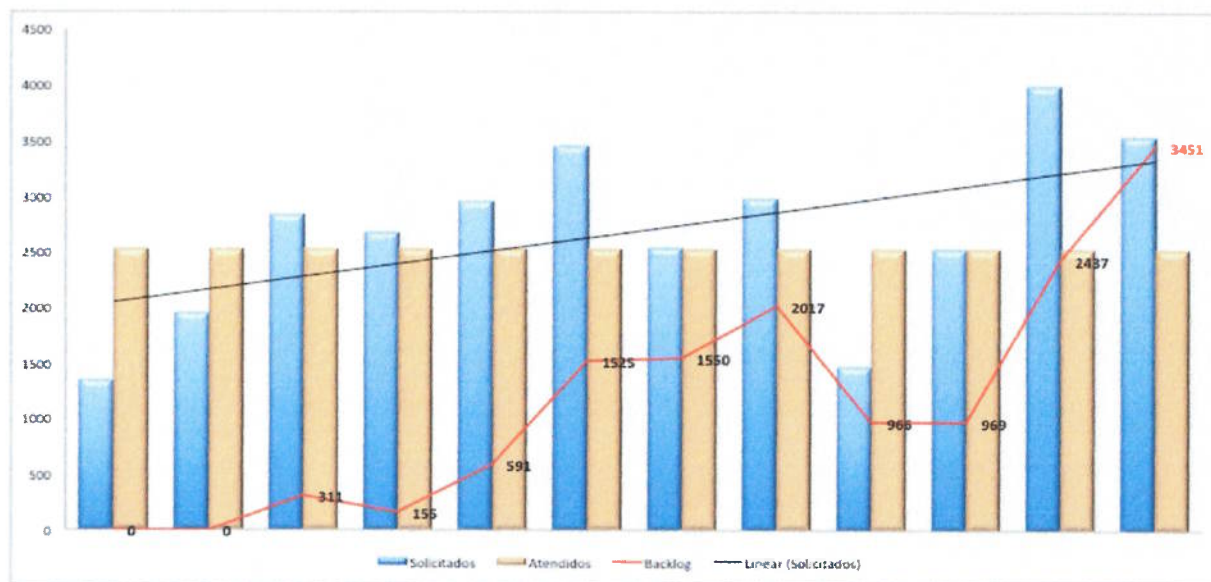
Porém, conforme já mencionado, apenas sistemas automatizados geram demandas constantes, logo, se o mesmo cálculo for realizado levando em consideração as variações da demanda ao longo dos meses o resultado é ainda maior. O *backlog* acumulado passa de 16.236 para 16.243 conforme apresentado no gráfico 3.

Gráfico 3 – Evolução do *Backlog* em 12 meses. Situação atual. Demanda variável

Sabendo que as horas disponíveis não podem ser estocadas, os dados obtidos na simulação da proposta final de recursos mostraram que apesar da projeção de *Backlog* ser significativamente menor, esse ainda persiste.

Com base nos dados de inclusão de novos servidores/analistas no sistema, o *backlog* ao final dos 12 meses resultou em 3.451 chamados em fila conforme disposto no gráfico 4.

Gráfico 4 - Evolução do *Backlog* em 12 meses. Situação proposta



De acordo com Taylor (2012) o estudo da incerteza deve ser baseado na análise completa dos períodos de demanda.

Se a demanda for constante ou variar positivamente, ou seja, se a demanda aumentar ao longo do tempo, é possível perceber que a capacidade produtiva não suportará essa evolução e que apesar de estar melhor administrável do que a situação atual por causa do menor volume, ao longo do tempo pode indicar perda no tempo de entrega dos serviços ao cliente e consequente insatisfação.

No entanto, se a demanda variar negativamente, ou seja diminuir ao longo do tempo, ou mesmo aumentar e diminuir de forma que ao no período o sistema permaneça em equilíbrio, a capacidade produtiva proposta conseguirá suprir e manter o *backlog* sob controle.

Se o sistema trabalhasse completamente otimizado, sem perdas, existiriam momentos de ociosidade que poderiam cobrir, conforme já mencionado um aumento de 6% da demanda.

Analísando a linha de tendência, é possível perceber que a demanda é positivamente variável, ou seja, que está crescendo e por isso existe uma forte tendência à formação de filas de espera. Analísando o valor do último mês analisado excede em 7% em relação a capacidade prevista e por isso, certamente os valores e variáveis precisarão ser ajustados para que o sistema volte ao equilíbrio.

7 CONCLUSÃO

Após a análise realizada foi possível utilizar a teoria das filas como aplicação prática para análise de um sistema produtivo de serviços.

Além disso, os cálculos dos tempos de fila e de congestionamento do sistema possibilitaram entender a principal causa para o gargalo na entrega do serviço analisado e ainda o início de um estudo de solução.

As simulações complementaram a análise a respeito das restrições de implementação da solução, porém não foram suficientes para a medição completa de ganho e retorno que a implementação da proposta pode trazer. Isso porque, para a análise completa do retorno, é necessário levar em consideração além da variabilidade da demanda, os custos envolvidos tanto de contratação, como para a entrega do serviço, e ainda o ganhos em satisfação, fidelização do cliente, renovação e indicação para novas oportunidades, que sem dúvida, compõem o retorno total da solução. Essas são absolutamente possíveis de serem calculadas mas não foram objeto desse trabalho.

Esse trabalho pode servir de base para o aprofundamento do tema apresentado tendo em vista a abordagem das variáveis do serviço, incluindo a execução de uma simulação real na produção.

Pode contribuir para um trabalho futuro que deve analisar, entre outras, a distribuição das chegadas ao longo do dia e com isso a formação dos turnos de trabalho, possíveis concentrações e períodos de maior pico, bem como ausências e períodos de descanso, a otimização incorporando novas atividades aos analistas de maior custo sem que essas interfiram no tempo de entrega ao cliente e ainda a ampliação do estudo para outras filas de atendimento menos estáveis.

REFERÊNCIAS

- BONDI, A.B. **Characteristics of scalability and their impact on performance....** New Jersey: WOSP, 2000. 203p.
- BUSSAB, W.O; MORETTIN P.A. **Estatística básica**. São Paulo: Editora Saraiva, 2011. 540p.
- CHASE, B. R. **Administração da produção para a vantagem competitiva**. São Paulo: Artmed, 2004. 724p.
- COGAN, S. **Gerenciando as percepções nas filas de espera: Para Aumentar o Nível de Satisfação dos Clientes**. Rio de Janeiro: Qualimark, 1998. 109p.
- DAVIS, M. M. et al. **Fundamentos da administração da produção**. 3.ed. São Paulo: Artmed, 1999. 852p.
- FITZSIMMONS, A.J; FITZSIMMONS, A.M. **Administração de Serviços**. 4.ed. São Paulo: Artmed, 2004. 564p.
- GIANEZI, I.G.N.; CORRÊA, H.L. **Administração estratégica de serviços: Operações para a satisfação do cliente**. São Paulo: Atlas, 1994. 240p.
- GONCALVES, J.E.L. **As Empresas são Grandes Coleções de Processos**. Revista Administração de Empresas, São Paulo, V. 40, n. 1, p. 06-19, 2000.
- HARGREAVES, L.; ZUANETTI, R.; LEE, R. et al. **Princípios básicos sobre administração da produção**. Rio de Janeiro: Senac Nacional, 2005. 112p.
- INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. São Paulo. Apresenta base de dados para as análises desenvolvidas na fundamentação. Disponível em: <<http://www.ibge.gov.br>>. Acesso em: 17, 18, 19 ago. 2013.
- INSTITUTO PARACLETO. São Paulo. Apresenta figura de um sistema de filas. Disponível em: <<http://institutoparaclete.org>>. Acesso em 22 nov. 2013.
- JOHNSTON, R.; CLARK, G. **Administração das operações de serviço**. São Paulo: Atlas, 2002.
- KREIN, J.D. **Tendências Recentes nas Relações de Emprego no Brasil**. 2007. 319p. Tese (Doutorado) – Universidade Estadual de Campinas, Campinas, 2007.
- MARCELINO, P.R.P. **Afinal, o que é terceirização? Busca de ferramentas de análise e de ação política**. Revista Pegada, São Paulo, V. 8 , n. 2, p. 55-71, 2007.
- MIGUEL, P.A.C. **Metodologia de pesquisa em engenharia...** 2.ed. Rio de Janeiro: Elsevier, 2012. 280p.

MORAIS, M.A.C. **Análise da Operacionalidade do terminal de produtos....** 68p. Monografia (Especialização Logística em Transporte) - Departamento de Engenharia de Produção, Universidade Federal do Espírito Santo, Programa de Especialização, Vitória, 2001.

OLIVEIRA, Marco A. **Terceirização: estruturas e processos em cheque nas empresas.** São Paulo: Nobel, 1994. 174 p.

REZENDE, D. A. **Engenharia de software e sistemas de informação.** Rio de Janeiro: Brasport, 1999. 292p.

RIFKIN, Jeremy. **O Fim dos Empregos: O Declínio Inevitável dos Níveis de Emprego.** São Paulo : Makron Books, 1995. 348p.

SEVERINO, Antonio Joaquim Cortez. **Metodologia do Trabalho Científico.** 21ed. São Paulo: Cortez Editora, 2000. 278p.

SLACK, N. et al. **Gerenciamento de operações e processos.** São Paulo, Artmed: 2006. 552p.

TAYLOR, J.R. **Introdução à análise de Erros: O estudo de incertezas em medições físicas.** São Paulo: BOOKMAN, 2012. 329 p.

TENÓRIO, F.G. **Tecnologia da informação transformando as organizações e o trabalho.** São Paulo: Editora FGV, 2007. 216p.

SCHMENNER, Roger W. **Administração de operações em serviços.** São Paulo: Futura, 1999. 424p.