

UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Planejamento de equipes de atendimento para o *call center* CEMIG–D
baseado em um algoritmo de filas generalizadas

ALEXANDRE HENRIQUE MARTINS

BELO HORIZONTE – MINAS GERAIS

Novembro/2016

ALEXANDRE HENRIQUE MARTINS

Planejamento de equipes de atendimento para o *call center* CEMIG–D
baseado em um algoritmo de filas generalizadas

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção, como parte dos requisitos para a obtenção do título de Mestre em Engenharia de Produção.

Orientador: Prof. Anderson Laécio Galindo Trindade

BELO HORIZONTE – MINAS GERAIS

Novembro/2016

RESUMO

As empresas de distribuição de energia brasileiras estão submetidas às regras estabelecidas pela união por meio da Agência Nacional de Energia Elétrica – ANEEL. Essa agência reguladora define, por meio de resoluções, a forma como os serviços prestados devem ser realizados com foco na qualidade e no atendimento permanente aos clientes dessas empresas. Assim, um dos aspectos permanentes de gestão, por parte da Cemig D, é relativo ao atendimento aos clientes do *call center*. Este canal visa ao atendimento, por meio do número telefônico 116, e está sujeito a parâmetros de qualidade de atendimento para que, ao menos, 85% das ligações sejam atendidas em até 30 segundos e ainda para que, nesse mesmo intervalo de tempo, somente 4% das ligações sejam abandonadas. Com isso, o planejamento das equipes de atendimento deve ser realizado com foco nos limites impostos pela agência nacional. Com esse intuito, desenvolveu-se um programa de dimensionamento de pontos de atendimento, por meio de conceitos de filas generalizadas – $G/G/s$ –, em que os parâmetros de entradas de ligações e de taxas de atendimento podem ser descritos como uma distribuição em probabilidade qualquer e com controle de chamadas abandonadas. Esse programa pode ser comparado com o formato atual de planejamento do *call center* da Cemig D, que utiliza a metodologia baseada em filas $M/M/s$ e em estimação dos parâmetros, via análise de predição para o volume de chamadas recebidas, e também pelo tempo médio de atendimento. Os resultados alcançados com a presente pesquisa demonstram que o sistema de dimensionamento via filas generalizadas apresentou-se mais eficaz, uma vez que a distribuição de atendentes, por intervalo a cada 30 minutos, foi mais balanceada que a solução adotada pela empresa, além de não violar os indicadores de qualidade de atendimento e de gerar maior economia financeira no planejamento simulado no primeiro semestre de 2016.

Palavras-chave: *Call center*, dimensionamento de equipes, teoria de filas generalizadas, distribuições de probabilidade.

ABSTRACT

The Brazilian power companies are subjected to the federal guidelines established by Agência Nacional de Energia Elétrica - ANEEL. By means of resolutions, this regulatory agency sets the procedures of how the services should be provided focusing on quality and the constant assistance to the clients of those companies. Thus, one of the resilient management issues concerning Cemig D is related to client attendance through call center. This channel is settled to provide assistance through the telephone number 116 and it is subjected to assistance quality parameters in order to answer at least 85% of the calls in up to 30 seconds and to keep the abandoned calls rate at the waiting period up to 4% of the calls. So the attendance teams planning should be done with the limits set by the national agency on focus. For that reason, a software concerning the scheduling of attendance points was developed using the concepts of General Distribution Queuing System – G/G/s – in which the parameters of income calls and the attendance rates can be described as a particular probability distribution, also controlling the abandoned calls rate. This software can be compared with the actual planning pattern of Cemig D Call Center which uses the methodology based on Multi-server Queuing Model – M/M/s - and on estimation of parameters by analysis of the number of received calls prediction and by the average attendance time. The achieved results with this research show that the system of scheduling by General Distribution Queuing revealed to be more efficient, once for each 30 minutes of interval, the distribution of attendees was more balanced than the solution adopted by the company. Furthermore, the attendance quality parameters were not violated and the solution would generate a larger finance economy on the first semester of 2016 planning simulation.

Keywords: call center, scheduling teams, queuing theory, probability distributions.

LISTA DE ABREVIATURAS

<i>ANEEL:</i>	Agência Nacional de Energia Elétrica
<i>CA:</i>	Chamada Atendida
<i>CAb:</i>	Chamada Abandonada
<i>Cemig D:</i>	Companhia Energética de Minas Gerais – Distribuição
<i>CO:</i>	Chamada Ocupada
<i>COf:</i>	Chamada Oferecida
<i>CR:</i>	Chamada Recebida
<i>CRM:</i>	<i>Customer Relationship Management</i> – Gestão do Relacionamento com o Cliente
<i>CTA:</i>	Central de Teleatendimento
<i>IAb:</i>	Indicador de Abandono
<i>ICO:</i>	Indicador de Chamadas Ocupadas
<i>IID:</i>	Independente e Identicamente Distribuída
<i>INS:</i>	Indicador de Nível de Serviço
<i>IVR:</i>	<i>Interactive Voice Response</i> – Resposta Interativa de Voz
<i>PA:</i>	Ponto de Atendimento
<i>PA/hora</i>	Pontos de atendimento contratados por hora
<i>RC/CR:</i>	Gerência de Controle do Relacionamento Comercial
<i>REN:</i>	Resolução Normativa
<i>SAC:</i>	Serviço de Atendimento ao consumidor
<i>SLA:</i>	<i>Service Level Agreement</i> – Acordo de Nível de Serviço
<i>TMA:</i>	Tempo Médio de Atendimento
<i>URA:</i>	Unidade de Resposta Audível
<i>TT:</i>	<i>Turbo Table</i>

LISTA DE SÍMBOLOS

χ^2 :	Qui-quadrado
X :	Variável Aleatória
$f(X)$:	Função de densidade de probabilidade de X
$E(X)$:	Esperança de X
$Var(X)$:	Variância de X
$P(\cdots)$:	Probabilidade
e :	Base neperiana - $e \approx 2,71 \dots$
$\exp(\cdots)$:	Exponencial
$N(\mu, \sigma^2)$:	Distribuição Normal com média μ e variância σ^2
$F(X)$:	Função de distribuição acumulada
$L(\hat{\theta} y)$:	Função de verossimilhança de y em relação ao parâmetro $\hat{\theta}$
$\log(\cdots)$:	Logaritmo
ρ :	Fator de ocupação de um sistema de filas
L_q :	Quantidade esperada de clientes na fila
L :	Quantidade média de clientes no sistema
W_q :	Tempo médio de espera na fila
W :	Tempo total de espera no sistema
C_a^2 :	Coefficiente de variação quadrático de chegadas de elementos na fila
C_s^2 :	Coefficiente de variação quadrático da taxa de serviço
$h(X)$:	Função de risco da variável X
$\phi(X)$:	Função de densidade da Normal padrão (0,1)
$\Phi(X)$:	Função de distribuição da Normal padrão (0,1)

LISTA DE FIGURAS

Figura 1- Box-plot para a distribuição de ligações recebidas nas segundas-feiras, às 11h30min, do ano 2015.....	10
Figura 2 – Representação gráfica da distribuição Triangular	19
Figura 3 – Volume de ligações recebidas por ano – 2012 a 2016.....	41
Figura 4 – Distribuição de chamadas por semana – 2012 a 2016.....	42
Figura 5 – Distribuição de chamadas por intervalo de horário – 2012 a 2016.....	43
Figura 6 - Distribuição dos tempos médios de atendimento entre 2012 e 2016.....	45
Figura 7 – Tempo médio de atendimento (TMA) por ano – 2012 a 2016	46
Figura 8 – INS por ano – 2012 a 2016	48
Figura 9 – IAb por ano – 2012 a 2016.....	49
Figura 10 – Predição para o volume de ligações recebidas entre 2016 e 2018	55
Figura 11 – Modelos de distribuição aplicados ao volume de chamadas atendidas entre 11h30min e 12h00min para as segundas-feiras de janeiro	60
Figura 12 - Modelos de distribuição aplicados ao Tempo Médio de Atendimento entre 11h30min e 12h00min para as segundas-feiras de janeiro	61
Figura 13 – Taxa de abandonos entre 11h30min e 12h00min para as segundas-feiras de janeiro ...	64
Figura 14 – Comparação entre as metodologias de dimensionamento – Dias úteis.....	69
Figura 15 - Comparação entre as metodologias de dimensionamento – Sábado.....	70
Figura 16 - Comparação entre as metodologias de dimensionamento – Domingo	71
Figura 17 – Dimensionamento para dias úteis – Janeiro.....	73
Figura 18 – Dimensionamento para o sábado – Maio.....	73
Figura 19 – Dimensionamento para domingo – Maio	74

LISTA DE TABELAS

Tabela 1- Variabilidade de ligações semanais por ano.	42
Tabela 2- Média diária de ligações recebidas por período.....	44
Tabela 3- Variabilidade do tempo médio de atendimento por ano.	44
Tabela 4- Média diária de ligações recebidas por período.....	46
Tabela 5 - Chamadas abandonadas em relação ao volume de chamadas recebidas	47
Tabela 6 – Chamadas abandonadas ≥ 30 segundos em relação ao volume de chamadas recebidas	47
Tabela 7 – INS máximos e mínimos – 2012 a 2016	49
Tabela 8 – IAb máximos e mínimos – 2012 a 2016	50
Tabela 9 - Predição de chamadas recebidas CEMIG 2016	55
Tabela 10 – Estimativas para a Taxa de entrada de ligações e Taxa de Serviço – Segunda-feira – entre 11h30min e 12h00min - Janeiro	61
Tabela 11 – Dimensionamento semanal por período do ano	72
Tabela 12 – Diferenças entre os dimensionamentos – por mês	72
Tabela 13 – Comparação entre os procedimentos para dimensionamento do call center Cemig D .	77
Tabela 14 – Custo total dos pontos de atendimento.....	77
Tabela 15 – Planejamento de equipes por dia útil e mês	78
Tabela 16 – Planejamento de equipes por dia de fim de semana e mês	78

LISTA DE ALGORITMOS

Algoritmo 1 – Dimensionamento Turbo Table.....	52
Algoritmo 2 – Estimação de parâmetros e dimensionamento Cemig D.....	54
Algoritmo 3– Número mínimo de atendentes para um $INS \geq 85\%$	57
Algoritmo 4 – Ponderação pelas chamadas abandonadas	63
Algoritmo 5 – Dimensionamento para os dias úteis	67

SUMÁRIO

1	Introdução.....	1
1.1	Objetivos	3
1.1.1	Objetivo geral	3
1.1.2	Objetivos específicos	3
1.2	Justificativa	4
1.3	Estrutura do Trabalho	4
2	O Marco Regulatório	6
2.1	Atendimento telefônico – parâmetros regulatórios e qualidade dos serviços	6
2.2	Período típico	9
2.3	Considerações Finais	10
3	Referencial Teórico	11
3.1	Modelos de probabilidade	12
3.1.1	Distribuição Exponencial.....	13
3.1.2	Distribuição Poisson	14
3.1.3	Distribuição Weibull.....	15
3.1.4	Distribuição Gama	15
3.1.5	Distribuição Log-normal.....	16
3.1.6	Distribuição Normal	17
3.1.7	Distribuição Binomial Negativa.....	18
3.1.8	Distribuição Triangular.....	18
3.2	Métodos de estimação de parâmetros	19
3.3	Teste de Aderência e Critérios de informação: ferramentas para a escolha do melhor modelo probabilístico	21
3.4	Métodos de predição.....	23
3.4.1	Alisamento exponencial simples.....	25
3.4.2	Método linear de Holt.....	26
3.4.3	Método da tendência amortecida	26
3.4.4	Método Holt-Winters para tendência e sazonalidade	27
3.4.5	Medidas de precisão	28
3.5	Estudo de filas	29
3.5.1	Filas M/M/s	30
3.5.2	Modelo de filas G/G/s.....	33
3.6	Probabilidades de Abandono	35
3.7	Considerações finais	39
4	O call center da CEMIG D	40
4.1	Caracterização das chamadas recebidas.....	40
4.2	Índices Regulatórios	48
4.3	Considerações Finais	50
5	Dimensionamento do Call Center CEMIG D	51

5.1	Método Atual.....	51
5.1.1	Estimação dos parâmetros da fila via métodos de predição	53
5.2	Modelo Proposto	56
5.3	Considerações Finais	65
6	Comparação entre os Métodos de Dimensionamento	66
6.1	Algoritmo para definição do dimensionamento	66
6.2	Comparação quantitativa entre os métodos de dimensionamento	68
6.3	Teste de eficácia do modelo de filas generalizadas: Resultados para o 1º semestre de 2016	74
6.4	Considerações finais.....	79
7	Conclusões	80
7.1	Sugestões para trabalhos futuros.....	81
8	Referências Bibliográficas.....	82
	Anexo.....	84

1 INTRODUÇÃO

O *call center* é um ambiente tecnológico de telefonia em que há interação entre atendente e cliente, ou entre oferta e recepção de demandas por um processo de comunicação. Neste cenário, a central de comunicação pode oferecer serviços, como os de telemarketing, serviços de atendimento ao cliente (SAC), suporte técnico ou atendimentos a emergências (Robbins, 2007).

No caso geral da interação entre atendentes de uma empresa e seus respectivos clientes, as ligações telefônicas chegam à central de atendimento por meio de um serviço público de atendimento telefônico. Após esse passo, as chamadas são recepcionadas por um sistema de resposta audível (ou IVR, em inglês), em que o cliente pode selecionar alguma opção de atendimento ou mesmo ser direcionado para uma fila de atendimento. No momento em que a ligação é redirecionada para um agente de telemarketing, este faz o acolhimento das demandas do cliente e as registra em um sistema de Gerenciamento do Relacionamento com o Cliente (CRM, em inglês).

Nesse processo, há a vinculação de medidas que compõem a estrutura e o padrão de qualidade desse atendimento telefônico. Cada *call center* tem uma tipologia de oferta de serviços, ou seja, as centrais podem funcionar somente em horários comerciais (5 dias por semana-8 horas por dia) ou em turnos mais extensos como 24 horas por dia, 7 dias por semana e com atendentes divididos em 4 turnos de 6 horas. Além disso, essas centrais de relacionamento contam com o Acordo de Nível de Serviço (ou SLA, em inglês), em que são definidos níveis de qualidade de atendimento, tempo médio de atendimento e processo de gestão da recepção de chamadas.

No caso de empresas de distribuição de energia, o serviço de atendimento a clientes é exigido pela Agência Nacional de Energia Elétrica - ANEEL, órgão fiscalizador e regulador do sistema elétrico brasileiro. Assim, toda empresa desse segmento é obrigada, conforme a Resolução Normativa ANEEL nº 414/2010, a manter o atendimento à solicitação de serviços 24 h por dia, todos os dias da semana em seus canais de atendimento, tais como o *call center*, a *internet*, a mensagem de celular, o aplicativo e as redes sociais, exceto no atendimento presencial que funciona em dias úteis no horário comercial.

Além de regular os padrões de atendimento, a ANEEL também regula os recursos financeiros que devem ser cobrados dos clientes em termos de tarifas. A cada ano, esses recursos para manter todos os serviços da distribuição de energia têm sido cada vez mais escassos. Portanto, em todas as áreas das empresas de energia elétrica, há constantemente a busca por processos otimizados de atendimento aos serviços regulados, assim como aos clientes da distribuição de energia.

Na Companhia Energética de Minas Gerais - Distribuição (Cemig D), o teleatendimento é realizado por uma empresa terceirizada, por meio do número 116. Contudo, todo controle e gestão dos padrões de qualidade do atendimento são feitos por equipe própria da Superintendência de Relacionamento Comercial. A equipe responsável por essa gestão pertence à Gerência de Controle do Relacionamento Comercial (RC/CR). Nessa gerência, é feita a estimativa do número de agentes necessários para atingir parâmetros determinados pela ANEEL.

Em relação à estrutura de *call center*, o método para compor o dimensionamento de equipes de atendimento tem característica determinística, pois não leva em consideração a variabilidade de eventos que podem influenciar na quantidade de ligações que são recebidas a cada período de 30 minutos por dia. Por isso, os métodos atuais são muito diretos e podem gerar um superdimensionamento ou subdimensionamento pela falta de gestão de incertezas. Tais incertezas podem ser definidas como a falta de padrão na entrada de ligações no *call center* durante a semana, em um dia ou nos intervalos de 30 minutos, além do controle de chamadas abandonadas e “rechamadas”.

Atualmente, para realizar o planejamento da equipe de *call center*, a CEMIG baseia-se em ferramentas de predição para estimar o volume de chamadas e de estimativa para o tempo médio de atendimento em um determinado período, a fim de compor o quadro de agentes necessários para atender a essa demanda. Assim, o dimensionamento das equipes é definido de acordo com esses critérios e implementado em um sistema de dimensionamento chamado *Turbo Table*, que utiliza conceitos de filas M/M/s, em que a entrada de clientes e o tempo de serviço seguem, irrestritamente, uma distribuição exponencial.

Neste trabalho, pretende-se desenvolver uma técnica de dimensionamento de equipes de *call center* baseada nos conceitos de filas generalizadas (G/G/s) na parametrização das variáveis de entrada de ligações e do tempo de atendimento ao qual o cliente é submetido. De acordo com Gautam (2012), esses parâmetros da fila podem seguir

qualquer distribuição de probabilidade conhecida e não necessariamente explicados conforme uma distribuição exponencial, como é o caso do tipo de filas M/M/s.

Assim, é possível obter estimativas mais precisas para os parâmetros de entrada, uma vez que estes são definidos como o modelo probabilístico mais adequado às características numéricas. Além disso, implementou-se, no programa de dimensionamento via filas G/G/s, formulações relativas à probabilidade de abandono de clientes na fila. Esse conceito foi adaptado do trabalho desenvolvido por Garnett, Mandelbaum e Reiman (2002), em que programam o controle de abandonos em uma fila capacitada com s atendentes.

Os resultados obtidos nessa pesquisa foram satisfatórios em termos de distribuição de pontos de atendimentos. Verifica-se que o dimensionamento via filas G/G/s com controle de chamadas abandonadas apresenta redução no volume de atendentes necessários para atender os níveis de qualidade propostos pelo regulador, em relação ao modelo atual de planejamento por filas M/M/s com estimação dos parâmetros de entrada via métodos de predição. Em termos quantitativos, no planejamento de equipes simulado para o primeiro semestre de 2016 houve uma redução de aproximadamente 5% em dias úteis, ao ser utilizado o modelo de filas generalizadas, em comparação ao procedimento atual.

1.1 Objetivos

Pretende-se desenvolver neste trabalho uma metodologia de dimensionamento de equipes de *call center* para o setor elétrico baseada em conceitos de filas generalizadas (G/G/s) com controle de chamadas abandonadas. Com isso, deseja-se reduzir os custos de planejamento de equipes de atendimento em centrais telefônicas e otimizar a disposição dos atendentes ao longo dos dias e horários de uma semana de trabalho.

1.1.1 Objetivo geral

Minimizar os custos do dimensionamento do *call center* CEMIG, baseando-se em conceitos de filas generalizadas G/G/s com controle de chamadas abandonadas.

1.1.2 Objetivos específicos

- Contribuir para melhor conceituação de uma gestão ótima de um *call center* no cenário de distribuição de energia elétrica no mercado brasileiro;

- Modelar os parâmetros que compreendem entrada e tempo de atendimento de uma fila generalizada;
- Analisar se a prática atual de gestão do *call center* gera superdimensionamento da central de atendimento e, por consequência, aumento dos custos operacionais da área;
- Implementar o método de gestão de filas por meio do modelo G/G/s com controle de chamadas abandonadas e verificar se é mais eficiente que o modelo utilizado pela CEMIG com a utilização do *Turbo Table*;
- Avaliar os métodos propostos e os impactos na gestão do *call center*;
- Analisar a eficácia dessa proposta em relação à prática atual de dimensionamento aplicada pela gestão do *call center* CEMIG.

1.2 Justificativa

Esta pesquisa se justifica, já que toda distribuidora de energia elétrica é obrigada a manter serviços de atendimento telefônico, e, com recursos escassos, essa solução matemática colaborará para a redução de tempo de atendimento, assim como otimizará a força de trabalho direcionada para o *call center* e para os custos com a operação. Esta pesquisa, também, propõe uma dinâmica mais precisa em relação à tomada de decisões quanto ao alcance das metas regulatórias de qualidade no atendimento telefônico.

1.3 Estrutura do Trabalho

O capítulo 2 apresenta o marco regulatório do setor elétrico. Nele, discutiremos sobre os principais itens relacionados aos atendimentos que devem ser prestados no *call center*, bem como sobre os parâmetros de qualidade e os indicadores básicos de gestão e controle da central de atendimento.

No capítulo 3 apresenta-se o referencial teórico utilizado nesta pesquisa. Nessa seção, abordam-se os modelos de probabilidade utilizados para modelar os parâmetros da fila, os métodos de estimação desses parâmetros, os testes de aderência e de confirmação dos modelos adotados. Além disso, são definidos os métodos de predição utilizados para estimar o volume de chamadas, os modelos de filas implementados nos métodos de dimensionamento abordados e as funções de probabilidade de abandono de clientes na fila.

No capítulo 4, há a análise descritiva das chamadas recebidas e dos tempos médios de atendimento do *call center* Cemig D, entre janeiro/2012 e agosto/2016, enquanto no capítulo 5 são desenvolvidos os algoritmos de dimensionamento do *call center* Cemig D. Assim, é detalhado o método atual de dimensionamento, por meio de modelos de predição da entrada de chamadas, média do tempo médio de atendimento e aplicação do *Turbo Table*. Por conseguinte, nessa mesma seção, é apresentado o modelo proposto de dimensionamento baseado em filas generalizadas com controle de chamadas abandonadas.

No capítulo 6, são apresentadas as comparações de eficácia entre os métodos apresentados. Inicialmente, é feita uma comparação entre a metodologia *Turbo Table*, com modelagem dos parâmetros da fila conforme uma função exponencial, e a metodologia baseada em filas G/G/s. Após esse processo, foi implementado o modelo atual de dimensionamento baseado na predição do parâmetro de entrada na fila e na estimativa do tempo médio de atendimento, o que é comparado com o método das filas generalizadas.

Por fim, no capítulo 7 são apresentadas as considerações finais desta pesquisa e as perspectivas para trabalhos futuros.

2 O MARCO REGULATÓRIO

O setor elétrico brasileiro é formado por empresas concessionárias ou permissionárias que têm autorização do governo federal para fornecimento de energia em todo o território nacional. No caso da Cemig D, por exemplo, essa empresa atende ao todo 774 municípios do estado de Minas Gerais e possui aproximadamente 8,2 milhões de clientes.

As empresas de distribuição de energia estão sujeitas à regulação específica que orienta e estabelece regras gerais para o fornecimento de energia elétrica. A Agência Nacional de Energia Elétrica (ANEEL) é a autarquia federal responsável por regular a geração, a transmissão, a distribuição e a comercialização da energia. Além disso, essa agência reguladora também fiscaliza as concessões, implementa políticas e diretrizes para atuação do setor, estabelece as tarifas de energias, entre outras atividades.

Assim, pela implementação de diretrizes que norteiam o fornecimento de energia, as empresas do setor elétrico brasileiro estão submetidas às regras da Resolução Normativa ANEEL nº 414/2010. Essa resolução tem o papel de estabelecer “condições gerais de fornecimento de energia elétrica de forma atualizada e consolidada”. Ou seja, todos os serviços e formas de execução das empresas desse setor são parametrizados por essa norma.

Com foco neste trabalho, tem-se, no capítulo XV dessa resolução, o conjunto de normas para o atendimento ao público. Com isso, obtêm-se todas as regras necessárias para fornecer atendimento aos clientes por meio do serviço de *call center*. Esse setor é responsável por auxiliar os clientes da distribuição, quando estes pleiteiam alguma informação, desejam registrar reclamações de consumo, solicitar serviços diversos e acusar situações de emergência e urgência, como falta de energia, por exemplo.

2.1 Atendimento telefônico – parâmetros regulatórios e qualidade dos serviços

Conforme o artigo 183 da Resolução Normativa – REN - nº 414/2010, a distribuidora de energia deve disponibilizar o atendimento telefônico aos seus clientes de acordo com os seguintes critérios:

- I. gratuidade para o solicitante, independentemente se a ligação provier de telefone fixo ou móvel;
- II. atendimento até o segundo toque de chamada;

- III. acesso em toda área de concessão ou permissão, incluindo os municípios atendidos a título precário¹, segundo a regulamentação; e
- IV. estar disponível todos os dias, 24 (vinte e quatro) horas por dia.

A ANEEL obriga que toda empresa com mais de 60 (sessenta) mil unidades consumidoras tenha uma Central de Teletendimento (*call center*) implementada. No caso da Cemig D (Companhia Energética de Minas Gerais – Distribuição), esse serviço é contratado junto à empresa AeC, localizada no bairro Bonfim, em Belo Horizonte. Essa empresa fornece toda a estrutura de atendimento, com pessoal treinado de acordo com os procedimentos padrões do setor elétrico, além de arcar com equipamentos e sistemas que permitem a fiscalização e o acompanhamento das ligações realizadas.

Em termos de fiscalização e monitoramento da qualidade do atendimento telefônico, todas as chamadas devem ser gravadas eletronicamente pelas distribuidoras, conforme artigo 187 da resolução normativa ANEEL nº 414/2010. Contudo, essas gravações devem ser realizadas com o prévio conhecimento dos respectivos interlocutores e fornecidas ao consumidor sempre que solicitadas.

Além disso, o órgão regulador impõe às distribuidoras de energia que trabalham com CTA – Central de Teletendimento – alguns indicadores que apuram a qualidade do atendimento aos clientes que são mensurados diariamente, mensalmente e anualmente, conforme o artigo 188 da resolução normativa nº 414/2010. Tais indicadores são o Indicados de Nível de Serviço (INS), Indicador de Abandono (IAb) e Indicador de Chamadas Ocupadas (ICO).

O INS é um indicador de qualidade que afere a razão entre o número de chamadas atendidas em até 30 segundos e as chamadas recebidas, excluindo-se as chamadas abandonadas em até 30 segundos. Esse índice é calculado por

$$INS = \frac{\sum_{i=1}^{365} CA_i \leq 30s}{\sum_{i=1}^{365} CR_i - \sum_{i=1}^{365} CAb_i \leq 30s} \times 100 \quad (1)$$

Em que:

$i = 1, \dots, 365$ representa os dias de um ano

CA_i = chamadas atendidas por dia de um ano

$CA_i \leq 30s$ = "chamadas atendidas cujo tempo de espera na fila é menor ou igual a 30 (trinta) segundos por dia

¹ Unidades Consumidoras localizadas em outras áreas de concessão ou permissão

$CAb_i \leq 30s$ = chamadas abandonadas cujo tempo de espera na fila é menor ou igual a 30 (trinta) segundos por dia

CR_i = chamadas recebidas por dia

O IAb é o índice que mede a razão das chamadas abandonadas com tempo de espera acima de 30 segundos em relação às chamadas recebidas, excluindo-se as chamadas abandonadas em até 30 segundos, medido por

$$IAb = \frac{\sum_{i=1}^{365} CAb_i > 30s}{\sum_{i=1}^{365} CR_i - \sum_{i=1}^{365} CAb_i \leq 30s} \times 100 \quad (2)$$

Em que:

$i = 1, \dots, 365$ representa os dias de um ano

$CAb_i > 30s$ = chamadas abandonadas com tempo de espera na fila maior que 30 (trinta) segundos por dia

$CAb_i \leq 30s$ = chamadas abandonadas com tempo de espera na fila menor ou igual a 30 (trinta) segundos por dia

CR_i = chamadas recebidas por dia

Por fim, o ICO representa a razão entre o número de chamadas ocupadas e as chamadas oferecidas e é calculado como

$$ICO = \frac{\sum_{i=1}^{365} CO_i}{\sum_{i=1}^{365} COf_i} \times 100 \quad (3)$$

Em que:

$i = 1, \dots, 365$ representa os dias de um ano

CO_i = chamadas ocupadas por dia

COf_i = chamadas oferecidas por dia

De acordo com o artigo 190, para empresas do porte da Cemig D, com mais de 60 (sessenta) mil unidades consumidoras, devem ser seguidas as metas para acompanhamento desses indicadores. O descumprimento desses referenciais pode ensejar em multas a serem pagas à ANEEL. Assim, os parâmetros para cada indicador são os seguintes:

- Indicador de Nível de Serviço: $INS \geq 85\%$;
- Indicador de Abandono: $IAb \leq 4\%$;
- Indicador de Chamadas Ocupadas: $ICO \leq 4\%$ até 2014 e $\leq 2\%$ a partir de 2015.

2.2 Período típico

Ainda segundo o artigo 188, entende-se que os indicadores diários são determinados pela utilização de registros observados em todos os períodos típicos entre 00h e 23h59min59s do dia em análise. Já o indicador anual é determinado pelos registros observados em todos os períodos típicos do ano em análise.

Um período típico é caracterizado por cada intervalo de 30 (trinta) minutos e recebe um volume de chamadas abaixo do limite proposto no artigo 188 da REN nº 414/2010. Nesse caso, utiliza-se a fórmula dos valores atípicos para estabelecer o limite de chamadas recebidas no *call center*, que tem como base a quantidade de ligações que entram na central no período anterior ao ano atual. Assim, os dados relativos aos mesmos períodos e dias da semana correspondentes ao ano anterior são organizados. A partir disso, procede-se com o seguinte cálculo:

$$\text{Limite} = Q3 + 1,5 \times (Q3 - Q1) \quad (4)$$

Em que:

Limite = valor limite de chamadas recebidas no período para fins de sua classificação como típico;

Q3 = Terceiro Quartil (Percentil 75) do volume de ligações para um determinado período

Q1 = Primeiro Quartil (Percentil 25) do volume de ligações para um determinado período

Por exemplo, seja a distribuições de ligações de uma segunda-feira, às 11h30min, do ano de 2015. Nesse período, o primeiro quartil é 1.128 ligações e o terceiro quartil de ligações recebidas é 1.307. Pela regra do período atípico, se em alguma segunda-feira às 11h30min, em 2016, o volume de ligações recebidas ultrapassar o limite de 1.575 ligações, então esse intervalo é expurgado e não é contabilizado para cálculo dos indicadores de qualidade do atendimento telefônico.

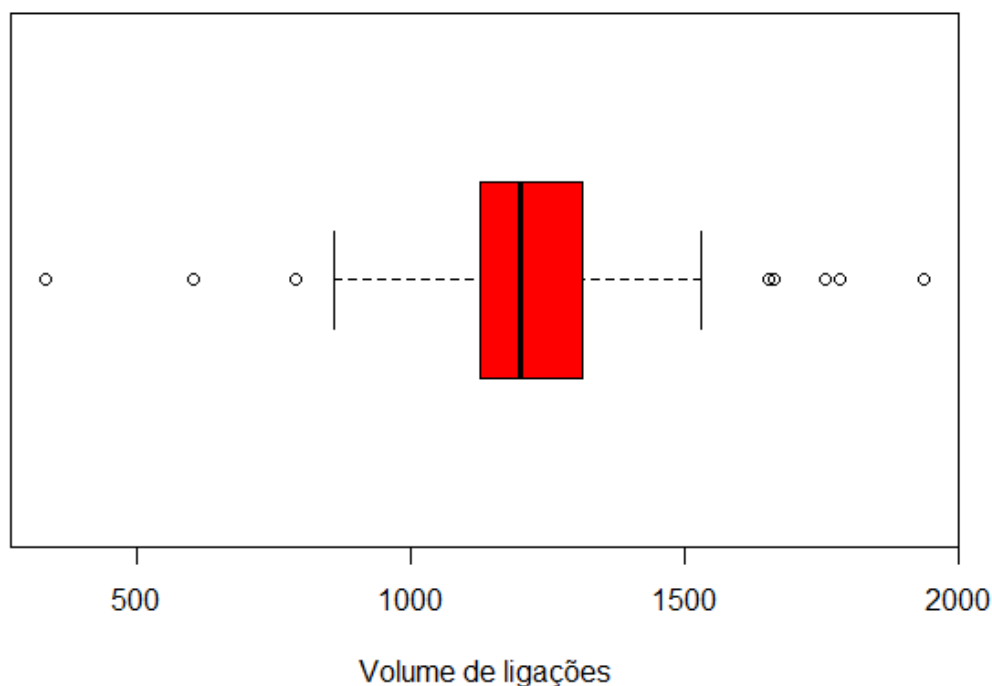


Figura 1- Box-plot para a distribuição de ligações recebidas nas segundas-feiras, às 11h30min, do ano 2015

2.3 Considerações Finais

Os limites regulatórios de qualidade do atendimento telefônico fazem com que as distribuidoras do setor elétrico definam as equipes necessárias para atender as metas estabelecidas. Qualquer mês de um ano em que os resultados propostos não sejam alcançados, haverá a incidência de multas e planos de ajuste de conduta a serem implementados pela distribuidora de energia.

Além disso, a ANEEL ainda permite que as concessionárias de energia tenham uma flexibilidade no atendimento, pois, considerando que todas elas estejam dimensionadas para atender os níveis do INS, do IAb e do ICO para um volume estimado de ligações, caso esses limites sejam ultrapassados, os períodos atípicos são expurgados para fins de cálculo desses indicadores. Contudo, o planejamento de equipes não considera a imprevisibilidade de um dia de atendimento no *call center* e faz a estimação dos pontos de atendimento para a projeção completa de ligações que podem ser recebidas.

Assim, os indicadores de qualidade telefônica devem ser utilizados para a construção da metodologia de planejamento de equipes baseada em teoria de filas. No caso das centrais telefônicas das distribuidoras de energia, a fila deve ser capacitada com um determinado número de servidores (ou atendentes) e ponderada pelos limites mínimos de qualidade de atendimento, como o INS e IAb.

3 REFERENCIAL TEÓRICO

Entende-se por filas sistemas que geram uma ordenação de elementos a partir das suas chegadas a determinado local para demandar algum tipo de atendimento. Esses elementos são sujeitos a um tempo de espera na fila e a um tempo de execução do serviço, o que determina a eficiência do sistema e a estrutura necessária para suportar o fluxo de atendimento.

Os sistemas de filas podem apresentar formas diversas quanto à regra de entrada de elementos e à taxa de execução do serviço. Os principais modelos abordados neste trabalho são o M/M/s e o G/G/s. No primeiro caso, tem-se que as entradas, que seguem um processo Poisson, e os tempos de serviço, têm distribuição exponencial independente e identicamente distribuídas. O modelo M/M/s é, dessa forma, capacitado com s atendentes. Diferentemente desse modelo, as filas G/G/s, também conhecidas como filas generalizadas, apresentam os parâmetros de entrada e tempo de atendimento baseados em distribuições quaisquer, ou seja, podem assumir o formato de diversos modelos de probabilidade, bem como carregam as características dessas distribuições.

Para definição dos parâmetros das filas generalizadas, observa-se que são necessárias as seguintes análises estatísticas para se definir a taxa de entrada e o tempo de serviço na fila:

- Definição do método mais eficiente para a modelagem dos dados baseados na máxima verossimilhança e no método dos momentos;
- Testes de aderência baseados no modelo χ^2 para avaliar se o modelo proposto é apto para explicar os dados relativos aos parâmetros da fila;
- Aplicação do critério de informação Akaike (AIC) para escolher o modelo de distribuição que melhor explique os parâmetros da fila;

Além dessas análises, trabalha-se em filas com conceitos de abandono. Ou seja, quando um elemento está na fila esperando por um atendimento e sua paciência ultrapassa um limite de tempo, então pode ocorrer o abandono da fila antes do atendimento ocorrer. Essa questão é trabalhada a partir da taxa de abandono e por meio de funções de probabilidades condicionais.

Uma alternativa à modelagem dos parâmetros é a utilização de predição baseada em dados de séries temporais. Dessa forma, definem-se também neste trabalho os conceitos de métodos estatísticos de predição baseados em modelos relativos a alisamento exponencial

para estimar os parâmetros de entrada, pois é o tipo de análise realizada atualmente pela equipe de planejamento do *call center*.

Assim, serão abordados neste referencial teórico os conceitos estatísticos necessários para a construção de uma análise de filas estatisticamente coerente. Nesta seção, as teorias de filas e de adequação de modelos de distribuição de probabilidade serão revisadas objetivando sustentar a construção do sistema de planejamento de *call center* baseado em filas generalizadas, tomando como exemplo a central telefônica da Cemig D.

3.1 Modelos de probabilidade

Neste trabalho, pretende-se definir um modelo de planejamento de equipes de *call center* baseado na gestão de filas. Em filas, há composição de parâmetros de entradas e de taxa de execução do serviço que são explicados de acordo com modelos de probabilidade.

De acordo com Montgomery e Runger (2002), probabilidade pode ser entendida como a chance de um determinado resultado ocorrer. Esses resultados podem ser associados a valores pertencentes ao conjunto do número reais e que são atribuídos aos resultados alcançados por essa probabilidade. Nesse caso, esses valores reais podem ter características discretas (valores contábeis) ou contínuos (valores inteiros ou relativos a medidas).

Assim, dado um conjunto numérico, discreto ou contínuo, associa-se uma função de probabilidade que define possíveis valores de uma variável aleatória X , por exemplo, e qual a chance da ocorrência do evento associado a essa variável. Essa função pode ser declarada como um modelo de probabilidade. Alguns modelos serão definidos nesta seção e têm caráter fundamental para a modelagem de parâmetros de uma fila.

As filas G/G/s e M/M/s têm definições e particularidades distintas entre si, apesar de pertencerem a um mesmo contexto científico. A fila G/G/s não convencionou um modelo específico para os parâmetros de entrada de elementos na fila e de execução do serviço. Já a M/M/s, declara explicitamente que esses parâmetros seguem uma distribuição exponencial, sendo que o parâmetro de entrada segue um Processo Poisson.

Com isso, foram escolhidos alguns modelos de probabilidade para auxiliar a modelagem de dados nesta pesquisa. Para os modelos de fila G/G/s, os dados relativos à entrada de ligações no *call center* Cemig D foram estudados conforme as distribuições Exponencial, Normal, Poisson, Gama e Binomial Negativa.

Para a avaliação da taxa de execução do serviço, as distribuições escolhidas foram a Exponencial, a Gama, a Log-normal, a Weibull e a Normal. Nesse caso, o tempo médio de atendimento no *call center* Cemig D pode ser entendido como o processo entre a captação da ligação (nascimento) e a finalização da chamada (morte), ou seja, característica suficiente para qualificar o TMA como um dado de tempo de vida. Já alguns intervalos podem se adequar com mais precisão à distribuição Normal, razão pela qual a referido distribuição fora estudada neste processo.

Portanto, nas próximas seções serão apresentadas distribuições em suas funções de densidade de probabilidade, esperanças e variâncias dos modelos. Assim, serão possíveis as análises dos parâmetros que compõem as filas e os cálculos das taxas.

3.1.1 Distribuição Exponencial

De acordo com DeGroot e Schervish (2011), essa distribuição é um caso especial da distribuição Gama e é definida como:

Seja o parâmetro $\beta > 0$. Uma variável aleatória X segue a distribuição Exponencial com parâmetro β , em que a função de densidade de probabilidade dessa variável é dada por

$$f_X(x) = \begin{cases} \beta e^{-\beta x} & , x > 0 \\ 0 & , x \leq 0 \end{cases} \quad (5)$$

Em que β é o parâmetro que expressa a taxa da distribuição e a esperança e a variância de X são dadas respectivamente por

$$E(X) = \frac{1}{\beta} \quad (6)$$

e

$$Var(X) = \frac{1}{\beta^2} \quad (7)$$

Uma propriedade importante dessa distribuição é que a Exponencial não guarda memória de cálculo, ou seja,

$$P(X \geq t + h | X \geq t) = P(X \geq h) \quad (8)$$

Em teoria de filas, essa distribuição é frequentemente utilizada também por apresentar essa propriedade, pois, a partir da falta de memória da Exponencial, define-se que os parâmetros tempo médio de atendimento e taxa de clientes que chegam à fila são

independentes para cada elemento que é atendido. Ou seja, o tempo médio de atendimento de um cliente que esteja na fila independe de TMA de um cliente em atendimento.

Além disso, Forbes *et al.* (2011), definem que a Exponencial pode ser entendida como a distribuição do tempo de ocorrência entre eventos, caso o número de ocorrências em qualquer intervalo tenha distribuição Poisson. A partir dessa definição, segundo DeGroot e Schervish (2012), há a relação da distribuição Exponencial com o processo Poisson, em que

“Suponha que a chegada de elementos em um determinado evento ocorra conforme um processo Poisson, cuja razão seja β . Seja Z_k o tempo até a k -ésima chegada com $k=1,2,\dots$. Se $Y_1 = Z_1$ e $Y_k = Z_k - Z_{k-1} \quad \forall k \geq 2$, então $Y_1, Y_2, \dots$ seguem uma distribuição Exponencial independente e identicamente distribuída com parâmetro β .”

Ainda na análise de filas, essa distribuição tem dupla utilidade na gestão desse elemento, pois auxilia na análise de chegada de clientes em um *call center*, bem como auxilia na medição do tempo médio de atendimento na central de atendimento, por meio da taxa de entrada de ligações e da taxa de execução do serviço, respectivamente.

3.1.2 Distribuição Poisson

Essa distribuição é aplicada na contagem de ocorrência de eventos (Forbes *et al.* (2011)). Por exemplo, pode-se modelar, por meio de uma Poisson, o volume de clientes que são recebidos em uma central telefônica.

Montgomery e Runger (2002) definem que, para certa quantidade de números reais, assume-se que as contagens ocorram entre os intervalos de tempo. Se esses intervalos puderem ser particionados em subintervalos pequenos o suficiente em que

1. A probabilidade de mais de uma contagem em um mesmo subintervalo é zero;
2. A probabilidade de contagem em um subintervalo é o mesmo para os outros subintervalos e proporcional ao tamanho do subintervalo;
3. A contagem de um subintervalo é independente dos outros intervalos (processo Poisson).

Assim, se uma variável aleatória X segue uma distribuição Poisson com parâmetro $\lambda > 0$ (*razão ou taxa*), a função de massa de probabilidade de X é dada por

$$f_X(x) = \frac{e^{-\lambda} \lambda^x}{x!}, x = 0, 1, 2, \dots \quad (9)$$

Em que:

$$E(X) = \lambda \quad (10)$$

$$Var(X) = \lambda \quad (11)$$

A Poisson pode ser utilizada para avaliação do volume de clientes que entram no *call center* em um modelo de filas G/G/s. Com essa distribuição, verifica-se a taxa de clientes que entram no sistema a cada intervalo de horário.

3.1.3 Distribuição Weibull

O tempo médio de atendimento de uma ligação em um *call center* pode ser entendido como o intervalo de tempo entre o nascimento e a morte da ligação de um cliente na central telefônica.

Neste caso, a distribuição Weibull é comumente relacionada a variáveis que medem o tempo de vida de produtos e de execução de processos. A função de densidade de probabilidade da Weibull, para uma variável aleatória X , é dada por:

$$f_X(x) = \frac{\beta}{\delta} \left(\frac{x}{\delta}\right)^{\beta-1} \exp\left[-\left(\frac{x}{\delta}\right)^\beta\right] \text{ para } x > 0 \quad (12)$$

Em que $\delta > 0$ é o parâmetro de escala e $\beta > 0$ é o parâmetro de forma, conforme definição de Montgomery e Runger (2002). Além disso, a esperança é dada por

$$E(X) = \delta \Gamma\left(1 + \frac{1}{\beta}\right) \quad (13)$$

e a variância dessa distribuição é definida como

$$Var(X) = \delta^2 \Gamma\left(1 + \frac{2}{\beta}\right) - \delta^2 \left[\Gamma\left(1 + \frac{1}{\beta}\right)\right]^2 \quad (14)$$

3.1.4 Distribuição Gama

Essa distribuição, cujos parâmetros são facilmente estimados pelo método dos momentos, tem como casos especiais as distribuições de Erlang e a Exponencial, de acordo com Forbes *et al.* (2011).

Conforme as definições de Montgomery e Runger (2002), a distribuição gama de uma variável aleatória X é dada por

$$f_X(x) = \frac{\lambda^r x^{r-1} e^{-\lambda x}}{\Gamma(r)}, \text{ para } x > 0 \quad (15)$$

em que os parâmetros de escala λ e de forma r são maiores que 0. Além disso, $\Gamma(r) = (r-1)\Gamma(r-1)$ é finito.

Para essa distribuição, tem-se que a esperança

$$E(X) = r/\lambda \quad (16)$$

e a variância

$$Var(X) = r/\lambda^2 \quad (17)$$

Para casos especiais, se r é um valor inteiro positivo, então $\Gamma(r) = (r-1)!$ e a distribuição Gama torna-se uma Erlang. No caso em que $r = 1$, então tem-se que a distribuição Gama transforma-se em uma Exponencial.

No programa de planejamento de *call center* desenvolvido, utiliza-se essa distribuição, tanto para analisar o volume de chamadas que chegam à central telefônica, a taxa de execução do serviço e as ligações abandonadas no *call center* Cemig D, quanto para analisar o tempo médio de atendimento.

3.1.5 Distribuição Log-normal

De acordo com Montgomery e Runger (2002), seja que um expoente de uma variável aleatória W é dado por $X = \exp(W)$, em que X é uma variável aleatória de interesse. Se W tem uma distribuição Normal, então X é chamada de distribuição Log-normal.

Assim, com W sendo uma distribuição Normal com média θ e a variância seja ω^2 , então, a variável aleatória X é Log-normal com função de densidade

$$f_X(x) = \frac{1}{x\omega\sqrt{2\pi}} \exp\left[-\frac{(\ln(x) - \theta)^2}{\omega^2}\right], 0 < x < \infty \quad (18)$$

em que,

$$E(X) = e^{\theta + \frac{\omega^2}{2}} \quad (19)$$

e

$$Var(X) = e^{2\theta + \omega^2}(e^{\omega^2} - 1) \quad (20)$$

Essa distribuição é frequentemente utilizada para modelar o tempo de vida de produtos que diminuem ao longo do tempo (Montgomery e Runger(2002)). Assim, a Log-

Normal pode ser utilizada no teste de modelagem para o tempo médio de atendimento do programa de planejamento de equipes do *call center* Cemig D.

3.1.6 Distribuição Normal

Essa é a distribuição mais frequentemente utilizada para modelar variáveis aleatórias. Conforme DeGroot e Schervish (2011), há três razões principais para que a distribuição seja utilizada:

1. Conveniência matemática, pois qualquer variável aleatória cujo modelo é Normal pode ter a sua função derivada explicitamente e terá uma forma simples;
2. Distribuição que pode ser adaptada para explicar diversos fenômenos físicos, mesmo que a variável aleatória de interesse sofra pequenas transformações;
3. Explicação, pelo teorema central do limite, de que, se uma amostra aleatória apresenta um volume razoavelmente grande de dados (pelo menos 30 unidades, por exemplo), é retirada de uma distribuição qualquer, cuja variância seja finita, a distribuição da média da amostra aleatória será Normal.

Segundo Montgomery e Runger (2002), se uma variável aleatória X tem distribuição Normal, então a sua função de densidade é dada por

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty \quad (21)$$

cujos parâmetros pertencem aos intervalos $-\infty < \mu < \infty$ e $\sigma > 0$. Além disso, tem-se que

$$E(X) = \mu \quad (22)$$

e

$$Var(X) = \sigma^2 \quad (23)$$

Para a Normal, a notação básica que representa essa distribuição é denotada por $N(\mu, \sigma^2)$.

Pelo fato de esta distribuição ser flexível para a modelagem de variáveis numéricas de naturezas distintas, aplicou-se, no programa de dimensionamento, a distribuição Normal para modelar os parâmetros de uma fila. Devido a isso, as entradas de clientes no *call center* e a taxa de atendimento foram ajustadas de acordo com um modelo de distribuição Normal e o resultado comparado com os outros modelos propostos neste capítulo.

3.1.7 Distribuição Binomial Negativa

De acordo com DeGroot e Schervish (2011), em vez de contar o número de sucessos em uma quantidade fixa de tentativas, pode ser necessário observar as tentativas até obter um número fixo de sucessos. Para isso, uma variável aleatória X que mede esse número fixo de sucessos segue uma distribuição Binomial Negativa.

Montgomery e Runger (2002) definem que, dada uma série de tentativas Bernoulli (tentativas independentes com probabilidade constante p de sucessos), seja variável aleatória X o número de tentativas até que r sucessos ocorram. Então X é uma variável aleatória Binomial Negativa com parâmetros $0 < p < 1$ e $r = 1, 2, 3, \dots$, cuja função de probabilidade é dada por

$$f_X(x) = \binom{x-1}{r-1} (1-p)^{x-r} p^r, \quad x = r, r+1, r+2, \dots \quad (24)$$

Em que

$$E(X) = \frac{r}{p} \quad (25)$$

e

$$Var(X) = \frac{r(1-p)}{p^2} \quad (26)$$

3.1.8 Distribuição Triangular

A distribuição Triangular é uma função que pode ser adaptada em estudos de dados contínuos e utilizada neste trabalho para realizar algumas simulações de chamadas abandonadas. Nesse caso, quando se deseja a geração de amostras aleatórias de bancos de dados em que é necessário evitar valores extremos, essa distribuição é utilizada para realizar essa operação.

Deste modelo, a função de densidade de distribuição é dada pela seguinte formulação para uma dada variável aleatória X

$$f_X(x) = \begin{cases} \frac{2(x-a)}{(b-a)(c-a)} & , se \ a \leq x \leq c \\ \frac{2(b-x)}{(b-a)(b-c)} & , se \ c \leq x \leq b \end{cases} \quad (27)$$

Em que

$$E(X) = \frac{(a + b + c)}{3}, \quad (28)$$

$$Var(x) = \frac{(a^2 + b^2 + c^2 - ab - ac - bc)}{18} \quad (29)$$

e

$$Moda = c \quad (30)$$

Nessa distribuição, tem-se que a é o coeficiente que representa o limite inferior (ou mínimo da distribuição dos dados), b é o limite superior (ou máximo da distribuição dos dados) e c é o parâmetro de forma (ou moda).

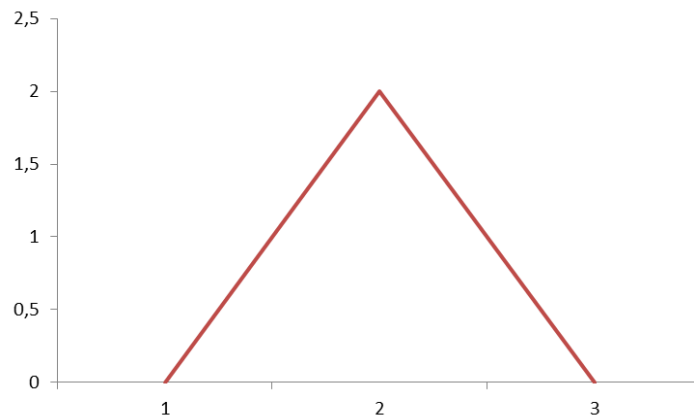


Figura 2 – Representação gráfica da distribuição Triangular

Para se trabalhar principalmente a simulação via distribuição Triangular, foi empregado o pacote do software R chamado “*triangle*”². Nesse pacote encontram-se funções para a distribuição Triangular e a função triangular de escala logarítmica para definição da moda entre os extremos.

3.2 Métodos de estimação de parâmetros

Após a definição das funções de distribuição utilizadas para a modelagem dos parâmetros de uma fila, necessita-se de técnicas apropriadas para a condução dos testes e de métodos para ajustar os dados amostrados pelo melhor modelo de probabilidade.

² É um pacote do software R que contém as funções de distribuição triangular e de uma distribuição triangular cuja escala é lognomal. As instruções de uso do pacote se encontram no sítio <https://cran.r-project.org/web/packages/triangle/triangle.pdf>.

Os métodos frequentemente utilizados nesse contexto são o da máxima verossimilhança e o dos momentos. Na análise de filas generalizadas, por exemplo, é imprescindível o ajuste dos parâmetros de entrada de ligações no *call center*, bem como o tempo médio de atendimento aos clientes pela função de probabilidade que melhor explique essas medidas. Com esses valores bem ajustados, o planejamento de equipes necessárias para prestar o atendimento aos clientes é realizado de acordo com a demanda e a capacidade de atendimento.

O método dos momentos visa ajustar o valor dos parâmetros de uma distribuição a partir das esperanças populacionais ponderadas pela esperança amostral, segundo Forbes *et al.* (2011). De acordo com Montgomery e Runger (2002), os momentos populacionais (ou esperanças populacionais) são funções dos parâmetros desconhecidos de uma função. Assim, essas funções são resolvidas com base nas informações amostrais, com o intuito de estimar dos parâmetros de uma distribuição.

Segundo DeGroot e Schervish (2011), o método dos momentos pode ser usado também para obter o valor inicial do estimador de máxima verossimilhança para a estimativa dos parâmetros de uma distribuição. Assim, para o desenvolvimento desse conceito, aplica-se o método de Newton, conforme definem esses autores:

“Assuma que $X_1, \dots, X_n$ formam uma amostra aleatória de uma distribuição que está indexada por um parâmetro k -dimensional θ e que tenha pelo menos k momentos infinitos. Para $j = 1, \dots, k$, seja $\mu_j(\theta) = E(x_1^j | \theta)$. Suponha que a função $\mu(\theta) = (\mu_1(\theta), \dots, \mu_k(\theta))$ seja uma função de θ . Seja $\mu(\mu_1, \dots, \mu_k)$ uma função inversa tal que, para todo θ ,

$$\theta = \mu(\mu_1(\theta), \dots, \mu_k(\theta)) \quad (31)$$

Assim, sejam os momentos amostrais definidos como $m_j = \frac{1}{n} \sum_{i=1}^n X_i^j$ para $j = 1, \dots, k$. Portanto, o método do estimado dos momentos de θ é dado por $\mu(m_1, \dots, m_j)$.”

O outro método de estimação dos parâmetros, o da máxima verossimilhança, foi desenvolvido pelo estatístico R.A. Fisher em 1920. O objetivo desse método é estimar o valor para que os parâmetros da distribuição maximizem a função de verossimilhança, como definem Montgomery e Runger (2002).

De acordo com DeGroot e Schervish (2011), sejam as variáveis aleatórias $X_1, \dots, X_n$. Assume-se que essas variáveis formam uma amostra aleatória de uma

distribuição contínua ou discreta qualquer em que a função de densidade seja dada por $f(x|\theta)$, em que θ pertence a algum espaço paramétrico Ω . Assim, θ pode ser ou um valor real ou um vetor. Para cada vetor $X = (X_1, \dots, X_n)$ na amostra, o valor da função conjunta de probabilidade será usualmente denotado por $f_n(x|\theta)$.

Então, de acordo como Montgomery e Runger (2002), a função de verossimilhança é dada por

$$L(\theta) = f(x_1|\theta).f(x_2|\theta). \dots .f(x_n|\theta) = \prod_{i=1}^n f(x_i|\theta) \quad (32)$$

Dessa forma, o estimador de máxima verossimilhança é o valor de θ que maximiza a função $L(\theta)$.

Esses dois métodos apresentados (método dos momentos e da máxima verossimilhança) são utilizados para a modelagem dos parâmetros das filas M/M/s e G/G/s. Segundo Gross *et al.* (2008), o método dos momentos executa rotinas de estimação mais rápidas que resultam nos parâmetros de uma distribuição, mas pode gerar valores diferentes do método da máxima verossimilhança.

Além disso, a máxima verossimilhança retorna estimativas mais consistentes do que o método dos momentos à medida que o tamanho da amostra aumenta. No entanto, conforme esses autores definem, o método dos momentos tem significativa importância para estudo de filas e, apesar de ser menos desejável que o método da máxima verossimilhança, o método dos momentos pode ser utilizado para esse tipo de análise.

3.3 Teste de Aderência e Critérios de informação: ferramentas para a escolha do melhor modelo probabilístico

Além do modelo específico para cada um dos parâmetros da fila, necessita-se realizar a análise de aderência dos dados amostrados em relação às funções de distribuição definidas na pesquisa e definir a melhor distribuição para aplicação. Para isso, utilizou-se o teste Qui-quadrado (ou χ^2), com nível de significância de 5% para validação do modelo, e a análise de critérios de informação para a escolha da melhor função de distribuição.

O teste de aderência faz uma comparação entre os dados observados em relação à expectativa de resultados de um determinado modelo. Taylor (1997) define como ação principal na análise de aderência de função de probabilidade a verificação da consistência dos valores observados em relação à distribuição teórica esperada.

De acordo com Rayner *et al.* (2009), esse teste de aderência também é chamado de Qui-quadrado (ou χ^2) de Pearson e é um dos primeiros testes de inferência, além de ser o mais utilizado em análises estatísticas. Ainda segundo esses autores, o teste χ^2 baseia-se na composição de uma hipótese nula, em que se afirma que uma amostra aleatória $X_1, \dots, X_n$ de tamanho n provém de uma população definida por um modelo de distribuição acumulada $F(x)$. Essa hipótese nula é contraposta a uma afirmação alternativa em que a distribuição que explica essa amostra é diferente de $F(x)$.

Estruturalmente, tem-se que as hipóteses formuladas estão representadas nas seguintes afirmações:

H_0 : A amostra $X_1, \dots, X_n$ é explicada pela distribuição $F(x)$

H_1 : A amostra $X_1, \dots, X_n$ é explicada por uma distribuição diferente de $F(x)$

Após a definição dessas hipóteses, deve-se estimar o valor da estatística χ^2 . Segundo Taylor (1997), se houver um experimento relativo a uma amostra aleatória da variável X em que haja informações suficientes para definir *a priori* a distribuição desses dados, essa variável é repetida N vezes e os possíveis resultados de X são divididos em n índices de $k = 1, \dots, n$, formando uma contagem de observações O_k . Em relação à distribuição esperada para X , forma-se a amostra de resultados esperados de X para cada índice k e tem-se os valores E_k .

Com isso, a estatística de teste para o cálculo do valor χ^2 é dada por

$$\chi^2 = \sum_{k=1}^n \frac{(O_k - E_k)^2}{E_k} \quad (33)$$

Assim, se o valor calculado de χ^2 for menor ou igual ao valor tabelado para a estatística qui-quadrado $\chi_{n-1, \alpha}^2$, em que $n - 1$ é o número de graus de liberdade e n representa o tamanho da amostra dos dados e α é o nível de significância do modelo, então a hipótese nula não é rejeitada. Conclui-se, portanto, que os dados amostrados seguem a distribuição de distribuição $F(x)$.

Com base nesse teste, para cada intervalo de atendimento no *call center* da Cemig D, é realizado no programa de planejamento de equipes o teste de aderência para os modelos propostos que podem explicar os parâmetros da fila. Portanto, se a função de distribuição não for aprovada no teste de aderência, então essa distribuição é descartada da análise e não poderá estimar a taxa de entrada, ou a taxa de serviço ou a taxa de abandonos, por exemplo.

Além disso, mesmo que uma variável analisada se enquadre para estimar os parâmetros da fila, é necessário avaliar qual é o melhor modelo que explica as medidas utilizadas no dimensionamento das equipes. Assim, utiliza-se a teoria de “critérios de informação” para selecionar o modelo que tem uma adaptação mais precisa aos dados amostrais em relação aos outros modelos implementados.

No programa de dimensionamento de equipes de *call center*, foi implementado o critério de informação de Akaike ou AIC, em inglês. Esse método de análise de modelos foi apresentado pelo estatístico japonês Hirotugu Akaike em 1973. No artigo de Akaike (1974), é defendido que o uso recorrente de modelagem estatística em vários campos da ciência requer que sejam estruturados modelos adequados para quantificar um conjunto finito de dados.

Diante disso, Akaike propõe a composição de uma abordagem prática dos modelos probabilísticos implementados, baseada nos resultados de máxima verossimilhança de cada modelo. De acordo com Akaike (1974), sob certas condições de regularidade, a medida de máxima verossimilhança é assintoticamente eficiente, o que torna a função de verossimilhança sensível a pequenas variações dos parâmetros em relação ao valor real da população.

Portanto, quanto menor o AIC, mais próximos os valores dos parâmetros da amostra estão dos resultados reais da distribuição proposta na análise. Assim, Burnham e Anderson (2004) definem matematicamente o critério de informação de Akaike como

$$AIC = -2 \log \left(L(\hat{\theta}|y) \right) + 2k \quad (34)$$

em que $\log \left(L(\hat{\theta}|y) \right)$ é o logaritmo natural da função de verossimilhança do modelo analisado e k é o número de parâmetros do modelo. O número de parâmetro da função de probabilidade escolhida para o teste tem a função de corrigir algum desvio (ou viés) de estimação do critério, principalmente quando se tem amostras pequenas.

3.4 Métodos de predição

Para se ter controle e previsibilidade de negócios, muitas áreas econômicas, tais como a indústria, as finanças e as prestadoras de serviços, necessitam de ferramentas que indiquem possíveis cenários em um estado futuro para as suas atividades. Por essa razão, frequentemente são utilizados conceitos e análises baseadas em predição via séries de tempo.

De acordo com Hyndman (2008), essas séries temporais são entendidas como um conjunto de dados que apresentam as seguintes características:

- Tendência: a direção/o sentido da série de dados;
- Sazonalidade: padrão observado (crescimento ou decrescimento) com uma determinada periodicidade esperada;
- Ciclo: é uma repetição que ocorre com certa regularidade, mas que é desconhecida e com mudança de periodicidade;
- Erro: é um fator aleatório e imprevisível de uma série.

Cada espaço de tempo, em uma série, é tratado como um período (ou espaço de estados), e a predição é realizada período por período (Bowerman e O'Connell, 1987). Com isso, a cada período predito, é realizada uma atualização na série para que o resultado a ser aferido no espaço de tempo posterior possa ser realizado com base nos parâmetros de um modelo (aditivo ou multiplicativo), que definirá a característica dos valores analisados.

Assim, conforme a notação proposta por Chatfield (2003), seja $x_1, x_2, \dots, x_N$ uma série temporal observada. O problema básico da análise de predição é estimar valores para x_{N+h} , em que $N + h$ indicam um espaço de tempo futuro e h é conhecido como horizonte de predição. De acordo com Chatfield (2003), há três tipos de métodos preditivos:

1. Subjetivo: em que a predição pode ser realizada tomando como base o julgamento, a intuição, os conhecimentos comerciais e qualquer outro tipo de informação relevante (Ex: experiência profissional);
2. Univariada: neste caso, a predição é realizada por um modelo ajustado com base nas informações presentes e passadas de uma série de dados. Ou seja, para estimar o valor de x_{N+h} , basta saber as informações dos dados $x_N, x_{N-1}, \dots$ e a partir deles ajustar um modelo linear simples de tendência global, por exemplo;
3. Multivariada: este tipo de predição é feita quando os dados da série histórica dependem parcialmente de uma série de valores auxiliares, ou seja, de variáveis exploratórias.

Neste capítulo, serão definidos os modelos de predição conhecidos como Alisamento Exponencial (ou suavização exponencial), com o objetivo de aplicação na modelagem de parâmetros de uma fila. De acordo com Chatfield (2003), essa é uma classe geral de procedimentos de predição que requer uma simples atualização dos valores estimados por um determinado modelo.

Para Bowerman e O'Connell (1987), esses métodos de modelagem atribuem pesos aos dados da série de forma desigual, isto é, valores mais recentes da série têm peso mais relevante na predição do que valores mais antigos. Assim, os valores mais atuais tendem a influenciar mais na estimação de valores futuros do que os dados mais remotos na série. Além disso, esse método de predição é mais indicado quando há pouca variação nos dados da série ao longo do tempo.

Além disso, muitas séries de dados apresentam tendência e sazonalidade, e essa característica dificulta a composição de um modelo mais robusto. Dessa forma é necessário retirar esses efeitos da série e torná-la estacionária, isto é, com valores que configuram uma variação constante ao longo dos períodos analisados.

Portanto, utilizam-se os métodos de Holt e Holt-Winters de predição. Conforme Chatfield (2003), o modelo Holt é utilizado para modelagem de séries que apresentam tendência, sem sazonalidade e possui dois parâmetros (nível e crescimento). Já o método Holt-Winters é capaz de estimar as curvas de predição cujos dados originais apresentam tendência e sazonalidade, além de possuir três parâmetros (nível, crescimento e sazonalidade). Nível e crescimento são medidas representativas de tendência.

3.4.1 *Alisamento exponencial simples*

De acordo com a notação utilizada por Hyndman (2008), seja um conjunto de dados observados até o espaço de estado $t - 1$ e deseja-se estimar o valor do próximo estado t y_t . Sendo o valor predito denotado por $\hat{y}_t$ e o valor observado de y_t está disponível, o erro de estimação é dado por $y_t - \hat{y}_t$. Assim, conforme Brown *apud* Hyndman (2008), um valor predito, ajustado pelo erro de predição é dado por

$$\hat{y}_{t+1} = \hat{y}_t + \alpha(y_t - \hat{y}_t) \quad (35)$$

em que α é uma constante (peso) entre 0 e 1.

Pode-se observar com isso que um valor predito depende do valor anterior estimado mais um fator envolvendo o erro de predição. Outra representação feita por Hyndman (2008) é definida como

$$\hat{y}_{t+1} = \alpha y_t + (1 - \alpha)\hat{y}_t \quad (36)$$

Com essa representação, interpreta-se que o valor predito de uma série histórica é a média do valor estimado mais recente e do mais antigo ajustado pelo peso α do modelo.

Assim, de acordo com Hyndman (2008), $\hat{y}_{t+1}$ representa a média móvel de todas as observações passadas com os pesos decrescendo exponencialmente.

Assim, essa função, em longo prazo, torna-se estacionária, pois o efeito dos pesos é menos significativo à medida que os períodos aumentam. Além disso, o alisamento exponencial simples é utilizado para bancos de dados que não apresentam tendência, sazonalidade ou quaisquer outras características que podem influenciar a análise de predição.

3.4.2 Método linear de Holt

Esse método de predição é uma extensão do alisamento exponencial simples que permite a estimação de valores futuros, mesmo que os dados analisados apresentem tendência. O modelo linear de Holt, de acordo com Hyndman (2008), possui três equações e duas constantes de alisamento (α e β^*), com valores entre 0 e 1, e pode ser escrito como

$$\text{Nível: } l_t = \alpha y_t + (1 - \alpha)(l_{t-1} + b_{t-1}) \quad (37)$$

$$\text{Crescimento: } b_t = \beta^*(l_t - l_{t-1}) + (1 - \beta^*)b_{t-1} \quad (38)$$

$$\text{Predição: } \hat{y}_{t+h|t} = l_t + b_t \quad (39)$$

Nesse modelo, l_t representa a estimativa de nível (valor em que a série de dados varia) e b_t é uma estimativa de crescimento ou decrescimento da série a cada espaço de estado (tempo) t . Conforme descrito por Hyndman (2008), b_t é uma média ponderada do crescimento prévio b_{t-1} mais o crescimento baseado na diferença de níveis sequenciais ($l_t - l_{t-1}$).

3.4.3 Método da tendência amortecida

Essa é uma modificação do método linear de Holt, descrito por Gardner *et. al.* apud Hyndman (2008). Nesse modelo é inserido um termo ϕ que permite suavizar (amortecer) a tendência na equação de predição, conforme a representação das funções 42 a 44:

$$\text{Nível: } l_t = \alpha y_t + (1 - \alpha)(l_{t-1} + \phi b_{t-1}) \quad (40)$$

$$\text{Crescimento: } b_t = \beta^*(l_t - l_{t-1}) + (1 - \beta^*)\phi b_{t-1} \quad (41)$$

$$\text{Predição: } \hat{y}_{t+h|t} = l_t + (\phi + \phi^2 + \dots + \phi^h)b_t \quad (42)$$

Assim, se $\phi = 1$, obtém-se o mesmo modelo linear de Holt. Se $0 < \phi < 1$, então a predição se aproxima assintoticamente de $\frac{l_t + \phi b_t}{(1-\phi)}$. Esse intervalo usualmente é adotado, pois, de acordo com Hyndman *et al.* (2008), se $\phi > 0$, evita-se que um coeficiente negativo aplicado seja menor que b_{t-1} e, se $\phi \leq 1$, evita-se que b_t tenha crescimento exponencial.

3.4.4 Método Holt-Winters para tendência e sazonalidade

Segundo Hyndman (2008), se uma série histórica não apresenta sazonalidade nem tendência, um alisamento exponencial simples é eficaz na estimação de valores futuros. Caso haja tendência linear, o método linear de Holt é o mais adequado. Contudo, quando há presença de tendência e sazonalidade, esses dois procedimentos não são indicados para prever os valores futuros da série.

Assim, para inserir o fator de sazonalidade na predição de valores de uma série histórica, foi desenvolvida por Charles C. Holt, em 1957, e estudada por Peter R. Winters, em 1960, uma metodologia que é conhecida atualmente como “Método Holt-Winters”. De acordo com Hyndman (2008), esse método é baseado em três equações: uma para o nível, outra para a tendência e uma terceira para a sazonalidade. Dependendo da forma como a sazonalidade é modelada, o método de Holt-Winters pode ter duas formulações: aditiva ou multiplicativa.

O modelo de sazonalidade multiplicativa é modelado como representação das funções 45 a 48:

$$\text{Nível: } l_t = \alpha \frac{y_t}{s_{t-m}} + (1 - \alpha)(l_{t-1} + b_{t-1}) \quad (43)$$

$$\text{Crescimento: } b_t = \beta^*(l_t - l_{t-1}) + (1 - \beta^*)b_{t-1} \quad (44)$$

$$\text{Sazonalidade: } s_t = \frac{\gamma y_t}{(l_{t-1} - b_{t-1})} + (1 - \gamma)s_{t-m} \quad (45)$$

$$\text{Predição: } \hat{y}_{t+h|t} = (l_t + b_t h)s_{t-m+h_m^+} \quad (46)$$

em que m é a quantidade de espaços de estado com sazonalidade. Além disso, $\hat{y}_{t+h|t}$ é a predição para h períodos à frente, e $h_m^+ = [(h - 1) \bmod m] + 1$. Os parâmetros $(\alpha, \beta^*$ e $\gamma)$ estão contidos no intervalo entre 0 e 1. De acordo com Hyndman (2008), como todos os métodos de alisamento exponencial, é necessário conhecer os valores iniciais dos

componentes e das estimativas dos parâmetros do modelo aferidos a partir da série histórica.

Já o modelo aditivo, apesar de não ser um modelo comumente utilizado, é estimado conforme as funções 49 a 52:

$$\text{Nível: } l_t = \alpha(y_t - s_{t-m}) + (1 - \alpha)(l_{t-1} + b_{t-1}) \quad (47)$$

$$\text{Crescimento: } b_t = \beta^*(l_t - l_{t-1}) + (1 - \beta^*)b_{t-1} \quad (48)$$

$$\text{Sazonalidade: } s_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1 - \gamma)s_{t-m} \quad (49)$$

$$\text{Predição: } \hat{y}_{t+h|t} = l_t + b_th + s_{t-m+h_m^+} \quad (50)$$

3.4.5 Medidas de precisão

Na análise de séries de tempo, é possível que mais de um modelo consiga ser ajustado para uma mesma base de dados. De acordo com Hoshmand (2010), há várias medidas de precisão que podem ser aplicadas em análise de predição. Essas medidas de avaliação desses modelos são ajustadas para determinar o erro médio que pode ser esperado ao longo do tempo.

Algumas dessas medidas podem ser representadas pelo desvio médio absoluto (MAD, em inglês) que mede o erro apresentado pela diferença entre os valores reais dos dados históricos e os valores preditos pelo modelo dividido pelo tamanho da amostra.

$$MAD = \frac{\sum_{i=1}^n |e_i|}{n} \quad (51)$$

em que $|e_i|$ é o módulo do erro observado da estimativa do modelo e n é o tamanho da amostra.

Além dessa medida, há também o erro médio quadrático (MSE), que pode ser calculado como

$$MSE = \frac{\sum_{i=1}^n e_i^2}{n} \quad (52)$$

De acordo com Makridakis *et al.* (1997), a vantagem do MAD é que essa medida é mais fácil de interpretar, enquanto a MSE tem um manuseio matemático mais simples. Outras medidas apresentadas são aquelas relativas ao erro percentual de precisão das estimativas ajustadas nos modelos de predição. Segundo esses autores, são aplicáveis nesse

contexto também o percentual de erro médio absoluto (MAPE) e o erro percentual médio (MPE). Assim, para Hoshmand (2010), essas medidas podem ser representadas como

$$MAPE = \sum_{i=1}^n \frac{|(e_i/Y_i) \times 100|}{n} \quad (53)$$

e

$$MPE = \sum_{i=1}^n \frac{(e_i/Y_i) \times 100}{n} \quad (54)$$

em que Y_i é o resultado atual dos dados amostrados, e_i é o erro associado à predição, ou seja, a diferença entre o valor atual e o valor predito.

Tem-se com isso que, quanto menor forem os resultados dessas medidas de precisão, melhor será o modelo escolhido para explicar os dados históricos.

3.5 Estudo de filas

De acordo com Groos *et al.* (2008), uma fila pode ser descrita como um sistema em que há a chegada de clientes procurando atendimento às suas necessidades e que, após um determinado intervalo de tempo de espera, têm seu serviço executado e é liberado do sistema. Hillier e Liebermann (2015) descrevem que a teoria de filas utiliza modelos que explicam vários sistemas em que há espera por atendimento. Assim, a fila é um elemento não material de organização para atendimento a demandas de forma organizada, por meio de noções de chegada e prioridade.

No caso do *call center* da Cemig D, o sistema que descreve a fila de atendimento é bem similar à definição de Groos *et al.* (2008). Os clientes que ligam solicitando atendimento da distribuidora de energia são direcionados para o atendimento eletrônico. Caso necessitem de atendimento pessoal, eles digitam a opção específica e são direcionados para um operador de telemarketing.

Na central de atendimento de empresas do setor elétrico, cada cliente pode esperar no máximo 60 segundos na fila. Além disso, 85% dos clientes na fila têm que ser atendidos em até 30 segundos e somente 4% desses clientes podem abandonar o sistema após 30 segundos de espera. Caso esses parâmetros de qualidade não sejam respeitados, a distribuidora pode ser punida com o pagamento de multas ao órgão regulador do setor elétrico.

Além disso, conforme Hillier e Lieberman (2015) definem, cada modelo de filas possui fórmulas que medem a performance de cada sistema. Essas medidas dizem respeito ao tempo médio total de espera na fila, à quantidade de clientes na fila e à taxa de ocupação, por exemplo. A partir dessas medidas, é possível realizar planejamento das equipes de atendimento com base em teoria de filas e adaptá-los aos níveis de qualidade estabelecidos pela ANEEL.

Portanto, descrevem-se nesta seção os dois modelos de filas que serão trabalhados nesta pesquisa: o modelo M/M/s, em que as taxas de chegada de elementos na fila e as razões de serviços realizados seguem a distribuição Exponencial e são capacitados com s servidores ou atendentes; e a fila G/G/s, que considera as entradas e as taxas de serviço na fila com distribuições quaisquer, ou seja, os parâmetros das filas podem assumir qualquer modelo probabilístico.

O foco deste trabalho é propor a solução via filas G/G/s, conforme definição de Gautam (2012), e compará-las com a formulação de filas M/M/s utilizada atualmente pela equipe de planejamento do *call center* da Cemig D. Portanto, é necessário fazer a explanação das características desses modelos para assegurar conceitualmente a adequação da formulação via filas generalizadas para dimensionamento de centrais de atendimento do setor elétrico.

3.5.1 Filas M/M/s

O modelo de filas M/M/s, de acordo com Hillier e Lieberman (2015), tem como definição principal que os parâmetros de entrada (a taxa de entrada segue um processo Poisson) de clientes no sistema a cada intervalo de tempo e a taxa de serviço seguem distribuições independentes e identicamente distribuídas, conforme o modelo Exponencial. Além disso, com fluxos grandes de elementos que entram na fila, no modelo M/M/s o sistema torna-se capacitado com pelo menos s atendentes.

Para cada intervalo, representa-se a taxa média de chegadas de clientes na fila como a constante λ e a taxa média de atendimento por um servidor como μ . Nessa abordagem, modelo apresentado é um caso especial do processo de nascimento e morte. De acordo com Gross *et al.* (2008), esse processo específico da cadeia de Markov de tempos contínuos³ é definido como

³ Em filas, trabalha-se com o conceito de nascimento e morte de elementos em um determinado sistema. Assim, a cadeia de Markov em tempos contínuos de acordo com Feldman e Valdez-Flores (2010) é descrita

“Seja o conjunto de estados $\{0,1,2,\dots\}$ tipicamente denotando a população de algum sistema. Quando o sistema está no estado $n \geq 0$, sendo n o número de pessoas que entram na fila, o tempo até a próxima chegada à fila (ou nascimento) é uma variável aleatória Exponencial com razão λ_n . Na chegada, o sistema passa do estado n para o estado $n + 1$ (ou nascimento). Quando o sistema está no estado $n \geq 1$, o tempo até a próxima saída (ou morte) é uma Exponencial com razão μ_n . No momento da saída, o sistema passa do estado n para o estado $n - 1$.”

Assim, tem-se que a taxa média de atendimento é dada por

$$\mu_n = n\mu \text{ para } n \leq s \quad (55)$$

$$\mu_n = s\mu \text{ para } n \geq s \quad (56)$$

Ou seja, se há menos clientes na fila do que atendentes disponíveis no sistema, a taxa de atendimento no n -ésimo estado é ponderada pelo número de pessoas que demandam atendimento. Já para o caso em que há mais elementos do que servidores na fila, a taxa de atendimento ocorre em função do número de atendentes disponíveis no sistema. Assim, os atendimentos são limitados à capacidade da fila.

Além disso, a medida da taxa de chegadas de clientes na fila é a constante λ para qualquer estado. Com isso, pode-se calcular o fator de utilização (ou ocupação) do sistema como

$$\rho = \frac{\lambda}{s\mu} \quad (57)$$

A partir dessa medida, define-se quando a fila está equilibrada ou não. Se $\rho \geq 1$, o volume de entradas na fila excederá a sua capacidade de atendimento. Caso contrário, tem-se que a capacidade de atendimento dos servidores disponíveis é maior do que quantidade de clientes que demandam atendimento.

Tem-se, nesse modelo de filas, a composição de algumas medidas de efetividade que aferem a qualidade da gestão do sistema e alguns atributos que auxiliam na otimização do sistema no atendimento às demandas recebidas. Assim, pode-se medir

1. A probabilidade de existirem exatamente n clientes na fila

como: Seja um processo $Y = \{Y_t; t \geq 0\}$ com estado de espaço finito E é um processo markoviano se a seguinte formulação é aceita para todo $j \in E$ e $t, s \geq 0$

$$P(Y_{t+s} = j | Y_u; u \leq t) = P(Y_{t+s} = j | Y_t)$$

$$P_n = \begin{cases} \frac{(\lambda/\mu)}{n!} P_0 & \text{se } 0 \leq n \leq s \\ \frac{(\lambda/\mu)^n}{s! s^{n-s}} P_0 & \text{se } n \geq s \end{cases} \quad (58)$$

em que

$$P_0 = \frac{1}{\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \times \frac{1}{1-\lambda/(s\mu)}} \quad (59)$$

sendo que P_0 é a probabilidade de não haver nenhum cliente na fila de atendimento.

2. A quantidade esperada de clientes na fila

$$L_q = \frac{P_0 (\lambda/\mu)^s \rho}{s! (1-\rho)^2} \quad (60)$$

3. A quantidade média de clientes no sistema (clientes na fila e em atendimento)

$$L = L_q + \frac{\lambda}{\mu} \quad (61)$$

4. O tempo médio de espera na fila

$$W_q = \frac{L_q}{\lambda} \quad (62)$$

5. O tempo total de espera no sistema

$$W = W_q + \frac{1}{\mu} \quad (63)$$

Outra medida utilizada na gestão de filas é a probabilidade de um cliente entrar no sistema e ter que esperar pelo atendimento. Esse valor é obtido por meio da função de probabilidade $W_q(t)$ que mede a chance de um cliente esperar t unidades de tempo na fila até que um atendente esteja disponível para atender às suas demandas.

De acordo com Gross *et al.* (2008), $W_q(0)$ é a probabilidade de um cliente não esperar tempo algum na fila para ser atendido, ou seja, ao entrar no sistema, ele é imediatamente atendido. Então, o complementar dessa chance, $1 - W_q(0)$ traduz a probabilidade de um cliente ter que esperar pelo menos uma unidade de tempo na fila até que seja atendido. Portanto, segundo Hillier e Lieberman (2015), $W_q(0)$ é descrita matematicamente como

$$P(W_q = 0) = \sum_{n=0}^{s-1} P_n = 1 - \frac{(\lambda/\mu)^s P_0}{s!(1-\rho)} \quad (64)$$

Dessa função de probabilidade, pode-se construir a probabilidade do tempo médio de espera na fila ser maior que um tempo t qualquer de tal maneira que

$$P(W_q > t) = (1 - P(W_q = 0)) e^{-s\mu(1-\rho)t} \quad (65)$$

Outra função considerada no estudo de filas M/M/s é a de probabilidade do tempo de espera na fila de atendimento, que é definida como

$$P(T_q > t) = \frac{(\frac{\lambda}{\mu})^s P_0}{s!(1-\rho)} e^{-(s\mu-\lambda)t} \quad (66)$$

Da equação 68, deriva-se a probabilidade condicional de espera da fila de atendimento, dado que há retenção de elementos na fila. Essa função é definida como

$$P(T_q > t | T_q > 0) = e^{-(s\mu-\lambda)t} \quad (67)$$

Assim, de acordo com Gross *et al.* (2008), a probabilidade de um cliente ter que esperar na fila um tempo $t \neq 0$ na fila pode também ser escrita como

$$C(s, \lambda/\mu) = 1 - P(W_q = 0) = \frac{\frac{(\lambda/\mu)^s}{s!(1-\rho)}}{\left(\frac{(\lambda/\mu)^s}{s!(1-\rho)} + \sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} \right)} \quad (68)$$

A equação 68 é também conhecida como Erlang-c e se baseia nas definições do modelo M/M/s. Conforme os conceitos definidos por Gross *et al.* (2008), esse modelo ignora algumas complexidades inerentes ao estudo de filas, como os abandonos e as “rechamadas” na gestão de um *call center*, por exemplo. Além disso, nesse modelo admite-se que a fila tenha capacidade infinita.

3.5.2 Modelo de filas G/G/s

As filas G/G/s, foco desta pesquisa, admitem que as entradas em qualquer sistema e os tempos de atendimento podem ser explicados por outras distribuições de probabilidade conhecidas. Isso quer dizer que, diferentemente do modelo de filas M/M/s, cujos parâmetros principais da fila têm distribuição Exponencial, os parâmetros de uma fila

G/G/s podem ser modelados com base nas outras distribuições de probabilidade conhecidas.

Conceitualmente, Gautam (2012) define que uma modelo de fila G/G/s é aquele em que a taxa de entrada dos elementos na fila seguem um sistema renovável com distribuição entre os intervalos de tempo cuja média é dada por $1/\lambda$ e coeficiente de variação quadrático de chegadas C_a^2 . Além dessa medida, tem-se que a taxa de atendimento segue também uma distribuição independente e identicamente distribuída com média $1/\mu$ e coeficiente de variação quadrático da taxa de serviço C_s^2 .

De acordo com DeGroot e Schervish (2011), o coeficiente de uma distribuição é a razão do desvio-padrão pela média dessa função. Essa medida representa a variabilidade dos dados de uma distribuição em relação à média. Nas medidas apresentadas por Gautam (2012), os coeficientes de variação quadráticos para os parâmetros de filas G/G/s são representados como

$$C_a^2 = \frac{Var[T_n]}{(E[T_n])^2} \quad (69)$$

e

$$C_s^2 = \frac{Var[S_n]}{(E[S_n])^2} \quad (70)$$

em que T_n representa a variável que define a taxa de entrada de clientes na fila a cada espaço de estado e S_n é a variável que mede a taxa de tempo de atendimento a cada espaço de estado. Tem-se também que as esperanças das distribuições dessas variáveis são dadas por

$$E[T_n] = \frac{1}{\lambda} \quad (71)$$

e

$$E[S_n] = \frac{1}{\mu} \quad (72)$$

Então,

$$C_a^2 = \lambda^2 Var[T_n] \quad (73)$$

e

$$C_s^2 = \mu^2 Var[S_n] \quad (74)$$

Além dessas definições, admite-se que um fila G/G/s siga o regime padrão do primeiro a chegar é o primeiro a sair (em inglês, *first-in-first-out*) e que a capacidade da fila seja infinita. Tem-se também que esse modelo possui no mínimo s atendentes disponíveis.

Em relação à aplicabilidade desse modelo de filas, é necessário estabelecer as medidas de eficiência que auxiliam na avaliação das condições de atendimento do sistema. Nesse caso, Gautam (2012) utiliza os resultados de uma fila M/M/s e faz ajustes para adequação ao formato das filas G/G/s. Essas medidas dizem respeito à quantidade de clientes na fila e ao tempo em que estes tendem a esperar até serem atendidos.

Com a formulação proposta por Gautam (2012), utiliza-se o cálculo do tempo médio de espera na fila (W_q), como

$$W_q = \frac{\rho^2 C_s^2 + C_a^2}{2 \lambda (1 - \rho)} \quad (75)$$

em que $\rho = \frac{\lambda}{s\mu}$ é a medida de intensidade de tráfego. Assim, por convenção, as outras medidas de performance são representadas como

- a) A quantidade média de clientes no sistema

$$L = \lambda W \quad (76)$$

- b) O volume esperado de clientes na fila

$$L_q = \lambda W_q \quad (77)$$

- c) Tempo total de espera no sistema

$$W = W_q + \frac{1}{\mu} \quad (78)$$

Como na fila M/M/s, para que um sistema operando como uma fila G/G/s esteja estável, é necessário que $\rho < 1$. Se $\rho > 1$, então a fila estará com sobrecarga de demanda e, por fim, caso $\rho = 1$, o sistema estará no limite do atendimento. Além dessas medidas, tem-se que a probabilidade de não existirem clientes na fila é dada pela relação $p_0 = 1 - \rho$.

3.6 Probabilidades de Abandono

Para definir a intensidade de chamadas abandonadas geradas em um *call center*, Garnett, Mandelbaum e Reiman (2002) constroem uma série de medidas de performance para analisar esse evento. Essas medidas propostas estão relacionadas ao tempo de espera

de um cliente na fila, à proporção de chamadas abandonadas, ao volume de atendentes ocupados e de clientes na fila.

Nessa perspectiva, utilizam-se algumas aproximações pela Normal e definições de probabilidade condicional para estimar esses valores baseando-se no modelo de filas que esses autores denominaram de Erlang-A. Esse modelo, segundo Mandelbaum e Zeltyn (2005), é uma junção das filas Erlang B⁴ e Erlang C⁵. Nesse caso, cada cliente tem uma curva de tempo de paciência, ou seja, dado um limite de tempo de espera na fila, o atendimento pode ser abandonado, caso o tempo oferecido de espera seja maior que a paciência do cliente para aguardar o atendimento.

De acordo com Whitt (2006), uma fila com padrões iguais ao de uma Erlang A segue um modelo M/M/s + M, sendo +M relativo aos tempos de abandono de chamadas. Esse modelo considera que as chamadas chegam à fila de acordo com um processo de Poisson, cuja taxa é igual a λ e os tempos de atendimento são independentes e identicamente distribuídos (IID), além de seguirem uma distribuição Exponencial cuja média é $\frac{1}{\mu}$. s atendentes trabalham paralelamente neste modelo e os tempos de abandono de atendimento seguem uma distribuição IID Exponencial cuja média $m = 1/\theta$. Esse regime de atendimento, ainda, segue o critério de “primeiro a chegar é o primeiro atendido”.

Assim, Garnett, Mandelbaum & Reiman (2002) definem os seguintes teoremas que norteiam a composição das medidas de performance do modelo com abandonos:

Teorema 1: Assuma que $\lim_{N \rightarrow \infty} \rho_N = \rho_\infty$, para $0 \leq \rho_\infty \leq \infty$ e $\rho_N = \lambda_N / N\mu$ como a intensidade de tráfego. Então, o comportamento limite da proporção de clientes que abandonam uma chamada é dado por

$$\lim_{N \rightarrow \infty} P_N\{Ab\} = \begin{cases} 0, & 0 \leq \rho_\infty \leq 1, \\ 1 - \frac{1}{\rho_\infty}, & \rho_\infty > 1 \end{cases} \quad (79)$$

Teorema 2: Assuma que

$$\lim_{N \rightarrow \infty} \sqrt{N}(1 - \rho_N) = \beta, \quad -\infty < \beta < \infty, \quad (80)$$

⁴ Erlang B é um modelo de filas em que se considera a existência de um abandono como a ocorrência de mais clientes na fila do que de atendentes no sistema.

⁵ Erlang C é um modelo de filas em que ocorrerá a espera para atendimento, caso não haja servidores disponíveis no sistema.

$$\lim_{N \rightarrow \infty} \theta_N = \theta, \quad 0 < \beta < \infty. \quad (81)$$

$Sq_N(0) \xrightarrow{d} q(0)$ e, então $q_N \xrightarrow{d} q$, em que q é a única solução da equação diferencial estocástica

$$\begin{cases} dq(t) = f(q)dt + \sqrt{2\mu}db(t), \\ f(X) = \begin{cases} -\mu(\beta + x), & x \leq 0, \\ -(\mu\beta + \theta x), & x > 0, \end{cases} \end{cases} \quad (82)$$

- 1) nesse caso, q_N é a sequência de um processo estocástico em que, ao centralizar o processo em volta do valor de N (número de atendentes), pode-se ter o valor absoluto do tamanho da fila ($q_N \geq 0$) ou o número de atendentes ocupados ($q_N \leq 0$);
- 2) $\beta = \frac{(N-R)}{\sqrt{R}}$ é uma constante dependente do nível de serviço desejado, N é o número de atendentes e $R = \frac{\lambda}{\mu}$ é a taxa de atendimentos por serviço executado;
- 3) b denota o Movimento Browniano padrão.

Teorema 3: Seja $v = [q/\mu]^+$, em que q resolve a equação diferencial estocástica no teorema 2. Então,

- 1) $\sqrt{N}v_N(t) \xrightarrow{d} v(t)$, $0 \leq t \leq 1$, em que ambos $v_N(\infty)$ e $v(\infty)$ são limites na distribuição, como $t \rightarrow \infty$, de $v_N(t)$ e $v(t)$, respectivamente;
- 2) $v(\infty)$ tem a função de distribuição F_v dada por

$$1 - F_v(x) = \begin{cases} w(-\beta, \sqrt{\mu/\theta}), & x = 0, \\ w(-\beta, \sqrt{\mu/\theta}) \cdot \frac{h(\beta\sqrt{\mu/\theta})}{\Psi(\beta\sqrt{\mu/\theta}, \sqrt{\mu\theta}x)}, & x > 0. \end{cases} \quad (83)$$

Em que

$$w(X, Y) = \left[1 + \frac{h(-xy)}{yh(x)} \right]^{-1}, \quad (84)$$

$$\Psi(X, Y) = \frac{\phi(x)}{1 - \Phi(x + y)}. \quad (85)$$

Além disso, $h(x)$ é a função de risco de uma Normal dada por $\frac{\phi(x)}{1 - \Phi(x)}$, $\phi(X)$ é a função de densidade de probabilidade de uma Normal padrão (0,1) e $\Phi(X)$ é a função de distribuição de probabilidade de uma Normal padrão (0,1).

Assim, Garnett, Mandelbaum e Reiman (2002) desenvolveram medidas de performance de uma fila com abandonos, em que T é a variável aleatória que representa o tempo de espera na fila e Ab é a variável aleatória que representa o abandono.

$$P(T > 0) \approx w(-\beta, \sqrt{\mu/\theta}) \quad (86)$$

que demonstra a probabilidade de um cliente ter que esperar na fila,

$$P\{T > t\} \approx w\left(-\beta, \sqrt{\frac{\mu}{\theta}}\right) \times \frac{h(\beta\sqrt{\mu/\theta})}{\Psi(\beta\sqrt{\mu/\theta}, \sqrt{N\mu\theta}t)} \times e^{-\theta t} \quad (87)$$

representa a probabilidade do tempo de espera ser maior do que t ,

$$P(Ab > 0) \approx \left[1 - \frac{h(\beta\sqrt{\mu/\theta})}{h(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)})}\right] \times w\left(-\beta, \sqrt{\frac{\mu}{\theta}}\right) \quad (88)$$

define a probabilidade de se abandonar uma chamada e as esperanças são representadas por

$$E[n^\circ \text{ de atendentes ocupados}] \approx \frac{\lambda}{\mu} - \left[1 - \frac{h(\beta\sqrt{\mu/\theta})}{h(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)})}\right] \times w\left(-\beta, \sqrt{\frac{\mu}{\theta}}\right) \times \frac{\lambda}{\mu}, \quad (89)$$

$$E[n^\circ \text{ de clientes na fila}] \approx \left[1 - \frac{h(\beta\sqrt{\mu/\theta})}{h(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)})}\right] \times w\left(-\beta, \sqrt{\frac{\mu}{\theta}}\right) \times \frac{\lambda}{\theta}, \quad (90)$$

e a probabilidade de abandono de chamadas, dado um certo tempo de espera é definida como

$$P\{Ab > 0 | T > t\} = 1 - \frac{\Psi(\beta\sqrt{\mu/\theta}, \sqrt{N\mu\theta}t)}{\Psi(\beta\sqrt{\mu/\theta} + \sqrt{\theta/N\mu}, \sqrt{N\mu\theta}t)} \cdot e^{\theta t}, t \geq 0. \quad (91)$$

Nos casos apresentados acima, θ é a medida que expressa a taxa de chamadas abandonadas durante a operação de um sistema de filas.

3.7 Considerações finais

Em problemas de filas generalizadas, a estimação dos parâmetros de entrada e de tempo de serviço deve ser trabalhada conforme as análises estatísticas de modelagem e avaliação de distribuições. Portanto, foram escolhidas para essa modelagem as distribuições que melhor explicam esse tipo de variável.

Diferentemente das filas M/M/s, o modelo G/G/s carece de mais análises para testar a aderência dos modelos aos dados que explicam os parâmetros desse tipo de fila. Além disso, a análise de variação dos parâmetros dos bancos de dados captados é também necessária para o cálculo das medidas de performance, que auxiliam na estimação das dimensões da fila de atendimento. Assim, é possível estruturar a fila de modo a otimizar os níveis de qualidade do atendimento proposto. Além disso, avaliar a taxa de abandonos também pode apresentar resultados interessantes, uma vez que a performance da fila pode ser aprimorada com o objetivo de redução de clientes que podem sair do sistema antes de serem atendidos.

Quanto aos modelos de predição definidos, esses são utilizados para a modelagem dos parâmetros de uma fila M/M/s para o dimensionamento das equipes da central telefônica da Cemig D. Assim, é possível comparar os métodos de dimensionamento por meio de filas M/M/s e G/G/s, em termos de redução de custos de contratação de equipes de *call center*.

Portanto, os conceitos trabalhados neste capítulo são de importância fundamental para a construção do algoritmo de planejamento de equipes do *call center* Cemig D, baseado em filas generalizadas.

4 O CALL CENTER DA CEMIG D

O *call center* Cemig D é uma estrutura de atendimento telefônico aos clientes capacitada para atender, de acordo com as regras regulatórias, em média, quase 30.000 ligações por dia. Para isso, essa empresa conta com uma equipe responsável pelo planejamento da estrutura de pontos de atendimento (ou servidores). Essa atividade é desempenhada pela Gerência de Controle do Relacionamento Comercial (RC/CR).

Assim, busca-se pelo dimensionamento das equipes do *call center* que:

- Os requisitos regulatórios para atendimento telefônico sejam cumpridos;
- Haja um equilíbrio entre performance de atendimento, quantidade de atendentes e recursos orçamentários disponíveis.

Portanto, serão trabalhadas neste capítulo as características da central telefônica da Cemig D em termos de volume de chamadas e de tempo médio de atendimento, que são os parâmetros implementados na análise de filas. Além disso, analisam-se nesta seção os resultados dos indicadores regulatórios de nível de serviço e de chamadas abandonadas. Com isso, busca-se demonstrar os valores que serão implementados na regra de planejamento de equipes por filas generalizadas e os limites regulatórios que devem ser observados.

4.1 Caracterização das chamadas recebidas

Nos últimos cinco anos (entre 2012 e 2016⁶), o *call center* recebeu em média 29.652 ligações/dia, em dias úteis, e aproximadamente 23.142 ligações/dia em finais de semana e feriados. Isso demonstra uma flutuação considerável entre os dias úteis e os fins de semana e impacta diretamente na composição de equipes necessárias para proceder aos atendimentos aos clientes.

Desde 2012, o volume de ligações vem caindo cerca 12% ao ano até 2014, em média. Em 2015, houve um aumento de aproximadamente 7,64% no volume de contatos em relação a 2014 (de 9.238.811 ligações em 2014 para 9.944.418 em 2015). No ano de 2016, tem-se que o volume de chamadas apuradas até o dia 11/08/2016 era de 5.241.740 e a projeção de ligações recebidas para este ano é de 8.595.897. Assim, haverá uma queda de 13,5% no volume de ligações em 2016 com relação a 2015. A evolução anual é apresentada na Figura 3.

⁶ Dados apurados até o dia 11/08/2016

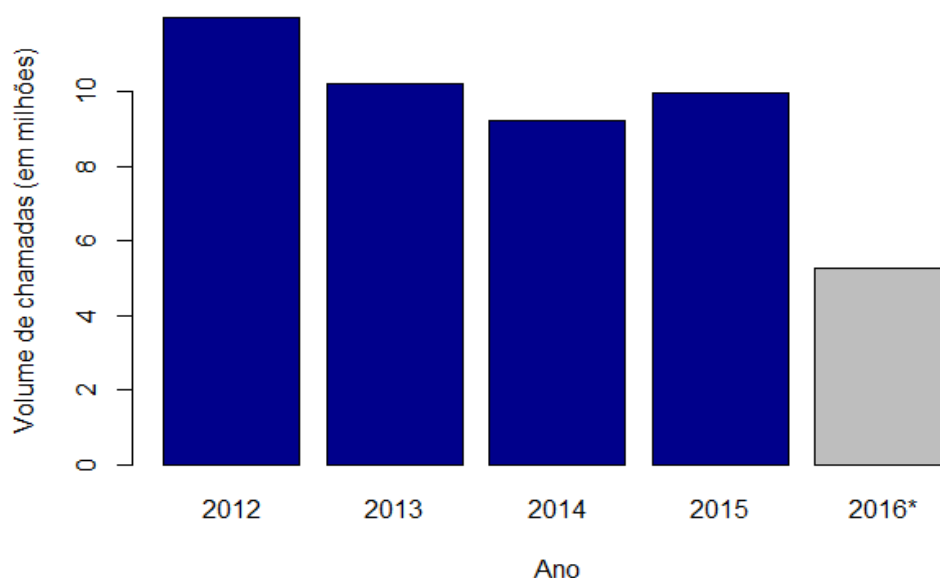


Figura 3 – Volume de ligações recebidas por ano – 2012 a 2016.

Fonte: RC/CR – Cemig D – 11/08/2016

Ao analisar a evolução do volume de chamadas recebidas semanalmente pelo *call center* CEMIG, observa-se também que a quantidade de ligações, em 2012, superou os outros anos analisados. Enquanto em 2012 foram recebidas em média 229.384 ligações por semana, em 2013 foram 194.898 ligações, em 2014 foram 175.798 ligações, em 2015 foram registradas 188.926 ligações. Pelos resultados apresentados até o dia 11/08/2016, tem-se que a média de ligações recebidas por semana até então foi de 159.144. Conforme a Tabela 1, pode-se avaliar também que as entradas de ligações por semana no *call center* é moderadamente homogênea (coeficiente de variação abaixo de 0,5). Assim, apesar de existir diferença considerável ano a ano, o volume de ligações semanais recebidas durante um ano é visivelmente parecido.

Tabela 1- Variabilidade de ligações semanais por ano.

Ano	Média – Vol. de Ligações Semanais	Desvio padrão Vol. Ligações Semanais	Coefficiente de Variação
2012	229.386	22.095	9.63
2013	194.898	32.207	16.53
2014	175.798	28.265	16.08
2015	188.927	33.581	17.77
2016	159.144	32.184	20.22

Fonte: RC/CR – Cemig D – 11/08/2016

Pela Figura 4, observa-se que a distribuição de ligações em 2016 é notoriamente inferior aos outros anos. Nesse ano, por exemplo, o volume de ligação em 75% da distribuição de chamadas recebidas por semana está abaixo de 177.400 ligações. Em 2015, em 50% das semanas o volume de chamadas recebidas foi inferior a esse valor. Essa análise corrobora com a análise de que há uma tendência de queda no volume de ligações entre os anos analisados. Assim, verifica-se que, em relação ao valor projetado de 2016, houve uma queda relativa de 28% no volume de ligações desde o ano de 2012.

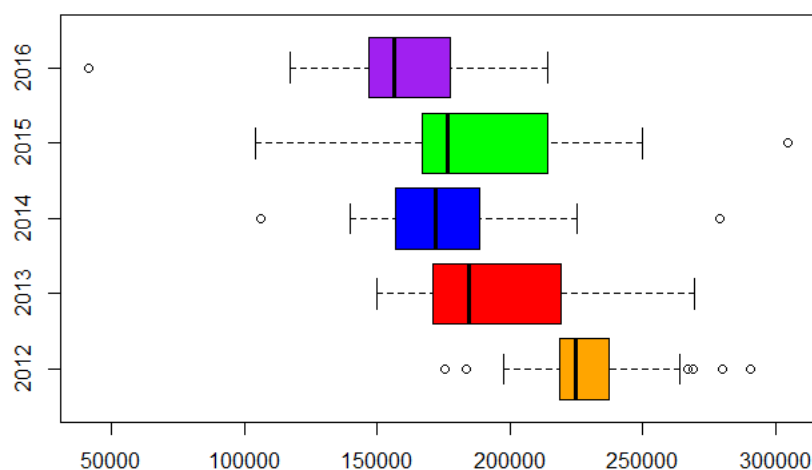


Figura 4 – Distribuição de chamadas por semana – 2012 a 2016.

Fonte: RC/CR – Cemig D – 11/08/2016

Em relação ao volume de chamadas recebidos por intervalo de horário, observa-se na Figura 5 a diferença de ligações atendidas no *call center* Cemig D nos últimos 5 anos. Nesse caso, o dia de atendimento na central telefônica foi dividido em 48 intervalos de 30 minutos. Ainda com base nesta figura, pode-se verificar um padrão comum no atendimento

independentemente do ano. Verifica-se que no período da madrugada (entre 00:00 e 05:30) o volume de ligação é extremamente baixo e a quantidade de chamadas recebidas varia entre 11 e 57 ligações. Entre 06:00 e 09:30, há um crescimento no volume de ligações. Entre 10:00 e 18:00, o volume de chamadas se estabiliza, mas mantém um volume significativamente alto (em média, 980 ligações a cada 30 minutos). Por fim, a partir das 18:30 o número de ligações começa a decrescer até chegar aos patamares do período da madrugada.

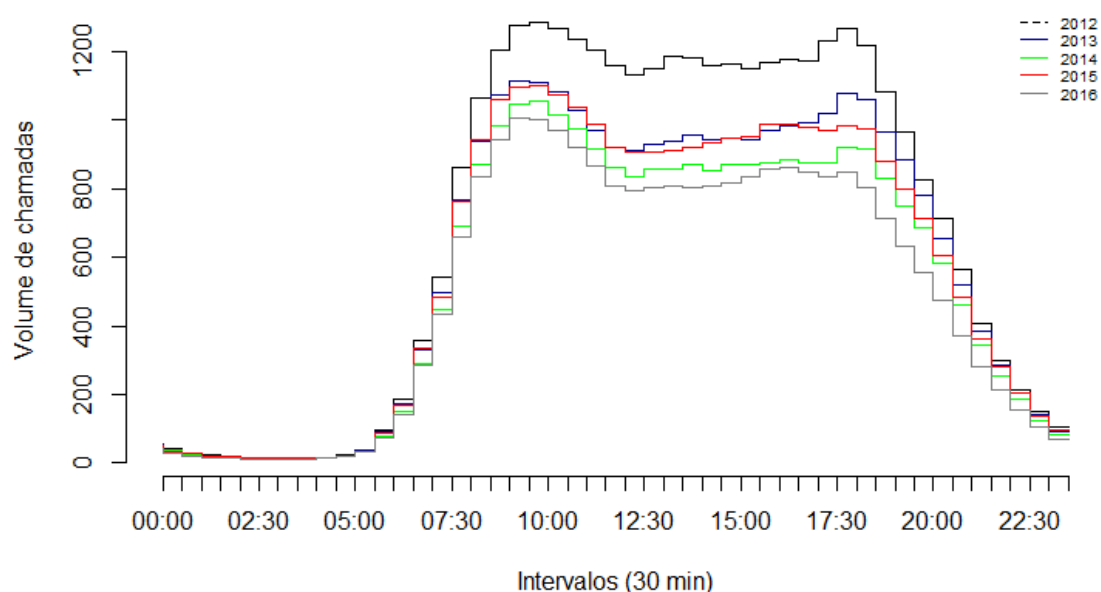


Figura 5 – Distribuição de chamadas por intervalo de horário – 2012 a 2016.

Fonte: RC/CR – Cemig D – 11/08/2016

Além dessas avaliações, deve-se considerar que o volume de ligações que entram no *call center* Cemig D pode ser diferente a depender do período do ano. Os meses entre janeiro e março e entre outubro e dezembro pertencem ao período úmido do ano, ou seja, é a parte do ano marcado por uma maior quantidade de precipitações pluviométricas. Já a outra parte do ano, entre abril e setembro, tem-se o período seco, que é quando a regra de chuvas não é frequente. Essa diferenciação é importante, pois, com a quantidade maior de chuva no 1º e no 4º trimestre de cada ano, há a maior possibilidade de ligações dos clientes para o *call center*, a fim de registrar ocorrências emergenciais, tais como falta de energia, fios partidos e quedas de árvores sobre a rede de distribuição.

Assim, pela Tabela 2, verifica-se que há diferença significativa entre o volume de chamadas recebidas no período úmido e seco. Além do aumento da diferença de chamadas recebidas nesses períodos ao longo dos anos (-7% em 2012 para -23% em 2016), observa-se também que o volume de chamadas recebidas por dia tem caído significativamente. Por exemplo, enquanto em 2012 recebiam-se, no período úmido, em média 33.772 ligações por dia e 31.649 no período seco, em 2016 esse volume caiu para 26.381 no período úmido e 21.436 no período seco. Esse evento é completamente justificável, pois nesses últimos 5 anos os clientes tiveram acesso aos canais virtuais (SMS, sítio eletrônico, redes sociais e aplicativos para *smartphone*) de atendimento. Enquanto em 2012 o atendimento virtual representava aproximadamente 9% das demandas dos clientes, em 2016 esse número está ultrapassando a barreira dos 30%.

Tabela 2- Média diária de ligações recebidas por período.

Ano	Média – Vol. de Ligações Diárias no Período Úmido	Média – Vol. de Ligações Diárias no Período Seco	% de redução de ligações no Período Seco em relação ao Período Úmido
2012	33.772	31.649	-7%
2013	29.918	25.998	-15%
2014	27.250	23.384	-17%
2015	29.725	24.779	-20%
2016	26.381	21.436	-23%

Fonte: RC/CR – Cemig D – 11/08/2016

Outro aspecto significativo na análise do perfil de atendimento no *call center* Cemig D é a evolução do tempo médio de atendimento ao longo dos anos. Devido à otimização nos procedimentos de atendimento e no sistema de registro de serviços, observa-se uma queda razoável no TMA apresentado na central de atendimento.

Tabela 3- Variabilidade do tempo médio de atendimento por ano.

Ano	Média – TMA	Desvio padrão TMA	Coefficiente de Variação
2012	288	16.65270	5.78%
2013	280	12.61557	4.51%
2014	291	16.72917	5.74%
2015	285	15.46223	5.43%
2016	271	15.77684	5.83%

Fonte: RC/CR – Cemig D – 11/08/2016

Pelos dados da Tabela 3, verifica-se que todos os tempos médios de atendimento são homogêneos e bem abaixo do parâmetro de 50% (os coeficientes de variação estão abaixo de 6%). Com isso, observa-se o TMA padrão para todos os atendentes do *call center*. Além disso, em 2016 tem-se o tempo médio de atendimento da série, com queda relativa de 6% em comparação com 2012 e 5% em relação ao ano de 2015. Observa-se também que a distribuição total dos tempos de atendimento tem média de 282,9 segundos, desvio padrão de 16,98 e coeficiente de variação de aproximadamente 6%, ou seja, os TMA são bastante homogêneos ao longo dos anos, conforme pode ser constatado na Figura 6.

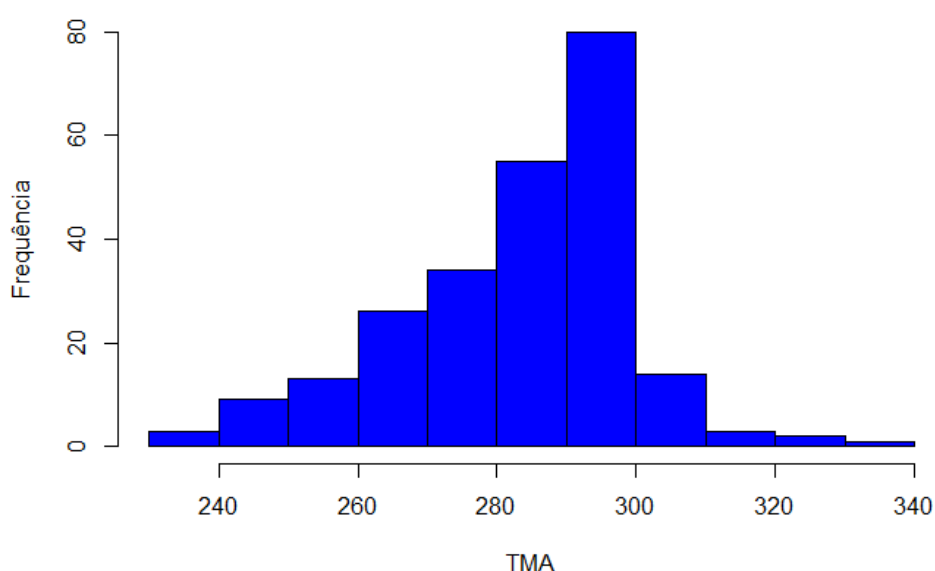


Figura 6 - Distribuição dos tempos médios de atendimento entre 2012 e 2016

Em relação à distribuição do TMA do *call center*, verifica-se que no ano de 2016 aconteceram os melhores rendimentos de atendimento também. Por exemplo, o maior tempo de atendimento verificado nesse ano foi de 291 segundos, igual à média de atendimento do ano 2014. Apesar de possuir valores extremos, o ano de 2012 apresentou menor variabilidade na distribuição dos tempos de atendimento. Já o ano em que se observou maior variabilidade na distribuição dos dados foi o ano 2014, em que houve a presença de muitos períodos com TMA extremos. Tais observações podem ser constatadas na Figura 7.

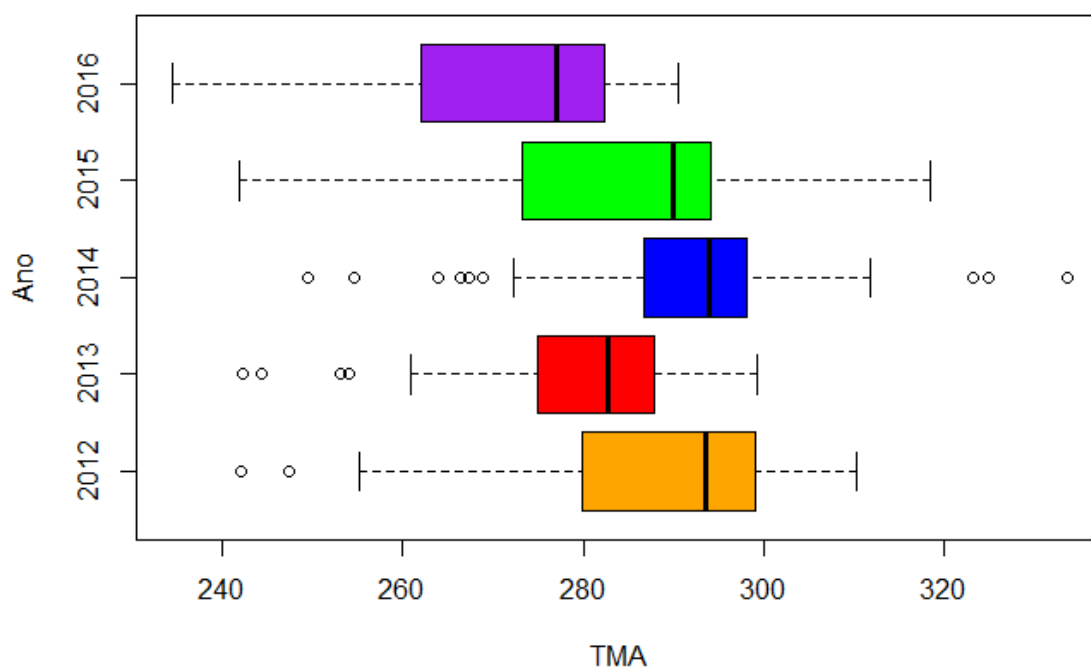


Figura 7 – Tempo médio de atendimento (TMA) por ano – 2012 a 2016

Assim como o volume de chamadas recebidas, o tempo médio de atendimento tende a ser diferente entre os períodos do ano. Pela Tabela 4, observa-se que, no período seco, o tempo médio de atendimento é menor do que no período úmido. Contudo, essa diferença não é progressiva ao longo dos anos e também não apresenta um decrescimento acentuado ao longo do período analisado. Isso é justificável, pois, como os atendentes passam por treinamentos a partir de procedimentos operacionais padrão, então, o TMA não é tão impactado ao longo dos anos.

Tabela 4- Média diária de ligações recebidas por período.

Ano	Média – TMA no Período Úmido	Média – TMA no Período Seco	% de redução do TMA no Período Seco em relação ao Período Úmido
2012	296,91	279,13	-6,37%
2013	281,39	278,27	-1,12%
2014	294,83	288,05	-2,35%
2015	284,90	284,28	-0,22%
2016	275,97	267,31	-3,24%

Fonte: RC/CR – Cemig D – 11/08/2016

Outro parâmetro relevante na análise de *call center* Cemig D é o volume de chamadas abandonadas. Pela Tabela 5, verifica-se que em 2012 foram registradas no total mais de 531.000 chamadas abandonadas para um volume de chamadas recebidas de 11.972.086. Esse abandono representou 4,44% de chamadas abandonadas. Em relação a essa proporção, observa-se que em 2013 e 2014 o volume de abandonos foi de 4,00% e 4,82% das chamadas recebidas, respectivamente. No ano de 2015, houve o maior percentual de chamadas abandonadas (7,60%). Já em 2016, foi apresentado o menor percentual para a série, em que 2,32% dos clientes que ligaram para o *call center* da Cemig D abandonaram a chamada.

Tabela 5 - Chamadas abandonadas em relação ao volume de chamadas recebidas

Ano	Total de Chamadas Recebidas	Total de Chamadas Abandonadas	% de Abandono
2012	11.972.086	531.858	4,44%
2013	10.202.689	408.597	4,00%
2014	9.238.811	445.111	4,82%
2015	9.944.418	756.054	7,60%
2016	5.251.740	122.091	2,32%

Ainda a respeito das chamadas abandonadas, avalia-se também o volume de abandonos após 30 segundos de espera na fila. Os níveis percentuais estão satisfatórios, uma vez que não foram retirados da base de dados analisada os intervalos atípicos. Apesar disso, nos anos de 2012 e de 2013 o volume de chamadas abandonadas ficou abaixo de 3% (2,84% e 2,80%), respectivamente. Em 2014, esse índice foi de 3,54%, o que também é satisfatório. Já em 2015, verifica-se o resultado maior da série com 5,59% de abandono. E em 2016, tem-se o melhor resultado apresentado pelo *call center* Cemig D nos últimos 5 anos com 1,54% de chamadas abandonadas.

Tabela 6 – Chamadas abandonadas ≥ 30 segundos em relação ao volume de chamadas recebidas

Ano	Total de Chamadas Recebidas	Total de Chamadas Abandonadas ≥ 30 segundos	% de Abandono
2012	11.972.086	339.820	2,84%
2013	10.202.689	285.897	2,80%
2014	9.238.811	327.424	3,54%
2015	9.944.418	556.059	5,59%
2016	5.251.740	80.959	1,54%

4.2 Índices Regulatórios

O Índice de Nível de Serviço (INS) do *call center* Cemig D demonstra o nível de qualidade do atendimento aos clientes da distribuição a partir da avaliação do volume de chamadas atendidas em até 30 segundos de espera na fila. Pelos limites estabelecidos pela ANEEL, a meta estabelecida para todas as concessionárias de energia elétrica é de pelo menos 85%. Pela Figura 8, tem-se o INS calculado desde o ano de 2012.

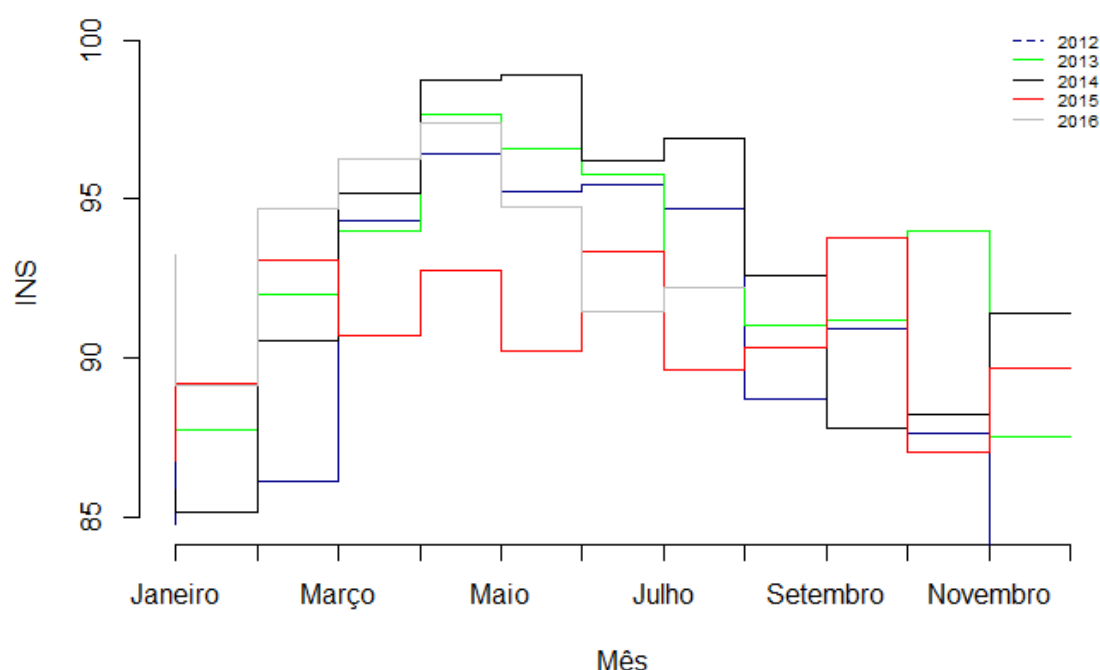


Figura 8 – INS por ano – 2012 a 2016

Observa-se, por meio dessa figura, que a Cemig D somente não alcançou a meta do INS nos meses de Janeiro/2012 (84,76%) e Novembro/2012 (80,61%). Nos outros anos avaliados, verificou-se que a central telefônica alcançou satisfatoriamente o INS mínimo de 85%. Destaca-se, neste período, conforme a Tabela 7, os valores máximos e mínimos de nível de serviço alcançados entre 2012 e 2016. Nesta tabela, tem-se que o planejamento do *call center* atingiu valores máximos nos últimos 5 anos entre 93,79% e 97,41%, extrapolando em pelo menos 11 pontos percentuais a meta estabelecida. Já os valores mínimos para esse indicador, excluindo-se 2012, ficaram entre 85,16% e 89,13%, ou seja, bem próximos do INS ANEEL.

Tabela 7 – INS máximos e mínimos – 2012 a 2016

Ano	INS máximo	INS mínimo
2012	96,41	80,61
2013	97,66	87,54
2014	98,89	85,16
2015	93,79	86,77
2016	97,41	89,13

O outro índice relativo ao *call center* e que é apurado pelas distribuidoras de energia é o Índice de Abandono (IAb). Esse indicador mede o número de chamadas abandonadas quando os clientes esperam mais de 30 segundos na fila até serem atendidos. Conforme estabelecido pela ANEEL, a meta para esse indicador é de 4% e, pelas informações da Figura 9, entre 2012 e 2016 esse indicador não foi violado pela Cemig D.

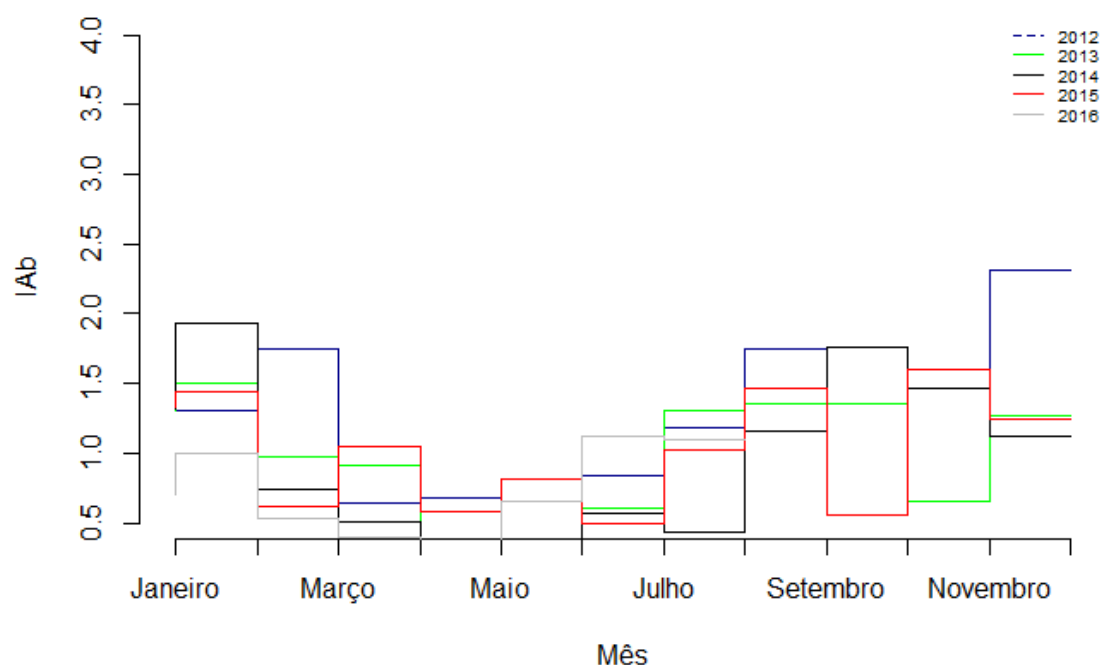


Figura 9 – IAb por ano – 2012 a 2016

Além disso, pela Tabela 8 identificam-se os valores para o indicador de abandono no período analisado. Nessa tabela, verifica-se que o valor máximo para o IAb está entre 1,12% e 2,31%. Já o mínimo para esse indicador é extremamente baixo e está entre 0,10% e 0,64%. Esse indicador demonstra que a qualidade do atendimento na central telefônica Cemig D está adequada aos parâmetros regulatórios.

Tabela 8 – IAb máximos e mínimos – 2012 a 2016

Ano	IAb máximo	IAb mínimo
2012	2,31	0,64
2013	1,50	0,34
2014	1,93	0,10
2015	1,59	0,50
2016	1,12	0,25

4.3 Considerações Finais

Pelos resultados apresentados, é evidente que o volume de ligações no *call center* Cemig D tem caído ao longo dos últimos 5 anos. Além disso, o TMA apresenta tendência de queda moderada para o atendimento aos clientes. Isso significa que, quanto menor o volume de chamadas e o TMA, menor deve ser a estrutura de atendimento disponibilizada na central de atendimento telefônico.

Além disso, percebeu-se também que os níveis de qualidade do atendimento têm *performance* que supera as metas regulatórias ($\geq 85\%$ para o INS e $\leq 4\%$ para o IAb). Ou seja, o INS tem facilmente ultrapassado a barreira dos 85%, sendo que em 2016 chegou ao máximo de 97,41%, e o IAb reduziu significativamente em relação à meta de 4%, uma vez que neste ano o mínimo valor foi de 0,25%.

Portanto, com esses resultados, pretende-se viabilizar uma solução baseada em teoria de filas que mantenha o equilíbrio entre a qualidade do atendimento e o custo para tal. Assim, o cenário ideal é aquele em que o dimensionamento das equipes do *call center* Cemig D atenda às metas regulatórias propostas pela ANEEL e que os custos com contratação de mão de obra sejam minimizados.

5 DIMENSIONAMENTO DO CALL CENTER CEMIG D

A Cemig D, por meio da gerência RC/CR, pertencente à Superintendência de Relacionamento Comercial, faz a gestão do *call center* que atende a demandas dos clientes da distribuição durante 24 horas por dia e 7 dias por semana. Nesse contexto, o dimensionamento é realizado para atender as chamadas no teleatendimento de acordo com os padrões regulatórios de qualidade do atendimento.

Para isso, atualmente, a equipe de planejamento da RC/CR executa o dimensionamento do *call center* baseando-se nas informações históricas do volume de chamadas que entram na central telefônica e no tempo médio de atendimento (TMA). Com essas informações, utiliza-se um macro do *Microsoft Excel Turbo Table* que executa o dimensionamento das equipes de acordo com conceitos de teoria de filas M/M/s.

Além dessa formulação, foi desenvolvida uma metodologia própria neste trabalho para contraposição à proposta do programa *Turbo Table*. Dessa forma, foi implementada uma estrutura de planejamento de equipes de *call center* baseada em filas generalizadas. Nas filas G/G/s, ou generalizadas, a taxa de entrada e a taxa de serviço baseada no TMA podem ter qualquer distribuição de probabilidade.

Assim, será apresentada nesta seção a formulação do dimensionamento via *Turbo Table* e o planejamento de equipes baseado em uma fila G/G/s. Além disso, em ambas as metodologias, o planejamento de equipes da central telefônica é feito para os 7 dias da semana e cada dia é separado em 48 intervalos de 30 minutos. Ou seja, nessa formulação cada intervalo representa uma fila e um dimensionamento distinto.

5.1 Método Atual

O *Turbo Table* pode ser encontrado no site <http://www.erlang.co.uk/> e foi desenvolvido na plataforma *Visual Basic for Applications* (VBA) para o *Microsoft Excel*. Essa ferramenta para *Excel* foi criada em 1999 e possui atualizações até o ano de 2001. Como um aplicativo de fácil utilização, o *Turbo Table* utiliza conceitos básico de filas M/M/s, em que tanto a entrada de ligações quanto a taxa de serviço são explicadas conforme uma distribuição Exponencial.

Com isso, há a necessidade de executar a modelagem dos parâmetros iniciais da fila que serão utilizados no dimensionamento. Assim, essa ferramenta foi reconstruída na

plataforma R^7 . Nesse contexto, foram utilizadas funções de modelagem estatística do R. A partir disso, o dimensionamento foi executado levando em consideração os conceitos das filas M/M/s e o nível de serviço mínimo regulatório de 85%.

Portanto, em termos de estrutura, o *Turbo Table* adaptado para a linguagem R é descrito conforme o algoritmo 1.

Algoritmo 1 – Dimensionamento Turbo Table

-
1. Gerar as entradas iniciais para as variáveis nível de serviço (SLA), TMA (AHT) e taxa de ligações por hora (*CallPerHour*)
 2. Calcular a intensidade de tráfego (*TrafficRate*)
 3. Calcular a taxa de Erlang/intervalo
 4. Estruturar o *call center* inicialmente com 100% dos agentes ocupados (nº de agentes)
 5. **Para** um número máximo de iterações (nº de agentes x 10) **faça**
 6. Calcular o número adequado de agentes para alcançar o SLA proposto
 7. Definir uma medida de precisão para o SLA e o número de atendentes
 8. **Fim Para**
 9. Exibir o número ideal de atendentes
-

O passo 1 do algoritmo 1 é gerado a partir das definições regulatórias do nível de serviço (mínimo de 85%) e pela modelagem probabilística dos parâmetros da fila (TMA e Taxa de entrada de ligações). Para isso, utilizou-se o método dos momentos para a estimação dos parâmetros da fila e as taxas de entrada e de serviço foram definidas a partir das esperanças estimadas das entradas das ligações a cada 30 minutos.

No passo 2, foram utilizadas as taxas de entrada de ligações e a taxa de tempo de serviço para calcular a intensidade do tráfego ou o fator de utilização (*TrafficRate*). Essa medida foi estimada com base na razão entre a taxa de entrada e a taxa do TMA. Essa medida avalia o equilíbrio do sistema e a capacidade de atendimento dos clientes que estão na fila. Assim, um sistema adequado é aquele que apresenta essa razão com valor abaixo de 1.

A taxa Erlang C calculada no 3º passo é também conhecida como a probabilidade de o cliente esperar na fila um tempo maior que 0 até que seja atendido. No caso de um

⁷ O R é um software livre de estatística computacional disponível em <https://www.r-project.org/>.

call center, essa medida traduz também a possibilidade de um cliente, ao solicitar o atendimento, ser direcionado para uma fila de espera.

Após a definição inicial do número de atendentes (passo 4), executa-se uma rotina de repetição para que, dado o valor pretendido do INS, o sistema retorne com o mínimo possível de atendentes necessários para que as ligações que entram no *call center* atendam o nível de serviço regulatório proposto. Assim, após atingir o número máximo de iterações, é exibida a quantidade de atendentes planejados para cada intervalo de 30 minutos a cada dia da semana.

5.1.1 Estimação dos parâmetros da fila via métodos de predição

Uma alternativa para estimar o parâmetro de entrada de ligações é a utilização de métodos de predição via análise de séries históricas. Esse tipo de abordagem é adotado pela equipe de planejamento do *call center* Cemig D para avaliar os valores relativos à entrada de ligações da central telefônica.

O procedimento utilizado para realizar a estimativa do volume de chamadas que entram no *call center* é desenvolvido para o ajuste ideal do número de atendentes para a central. Nesse caso, o dimensionamento estimado para cada período do ano, dia e intervalo de horário deve atender:

- O nível de serviço no atendimento aos clientes regulado pela ANEEL – Agência Nacional de Energia Elétrica;
- O aumento no volume de chamadas em determinados períodos do ano (exemplo, período chuvoso);
- Redução do número de atendimentos realizados no período seco;
- A redução constante do número de chamadas abandonadas.

Nessa abordagem, o tempo médio de atendimento pode ser estimado simplesmente pelo método de estimação de modelos de probabilidade. Assim, a taxa de atendimentos realizados a cada unidade de tempo pode ser mais precisa quando se calcula esse parâmetro por meio de uma média simples em um determinado período. Contudo, o processo utilizado pela equipe de planejamento da RC/CR é calcular o TMA pela média aritmética do período analisado e adotar esse resultado como parâmetro para o dimensionamento das equipes do *call center*.

Portanto, o procedimento para o dimensionamento das equipes do *call center* Cemig D é definido conforme o algoritmo 2.

Algoritmo 2 – Estimação de parâmetros e dimensionamento Cemig D

-
1. Gerar o modelo de predição para estimar o parâmetro de chamadas recebidas no *call center* Cemig D para o mês de planejamento desejado
 2. Ajustar o volume de chamadas recebidas por dia e intervalo do mês determinado
 3. Estimar a taxa de entrada de ligações por mês de dimensionamento, dia e intervalo de horário
 4. Calcular a taxa de serviço (ou TMA) por dia e intervalo do mês determinado
 5. Definir os parâmetros da fila e aplicar o algoritmo 1
-

Para gerar o modelo de predição, utiliza-se o volume de chamadas recebidas por mês de cada ano de uma série. Nessa perspectiva, a equipe de planejamento de equipes executa a análise de predição com base no histórico de ligações a partir de 2005 e aplica as funções de modelagem dos dados dessa série na plataforma *SAS Forecasting*⁸. Nesse programa é possível criar um projeto que analisa uma série histórica e retorna o modelo ótimo de predição com o menor erro de estimativa.

Por exemplo, no caso do volume de ligações recebidas pela CEMIG, verificou-se que o melhor modelo preditivo é o Método Aditivo de Winters – Aditivo. Nesse caso, foi aplicado para uma série de dados históricos compreendidos entre janeiro/2005 e dezembro/2015. O método de estimação utilizado é adequado para essa análise, pois é sensível a séries históricas que apresentam tendência e sazonalidade. Assim, há uma sazonalidade pertinente aos períodos do ano (úmido e seco) e uma tendência de crescimento no volume de chamadas nos meses compreendidos nos 1º e 4º trimestres do ano, além de uma redução entre os 2º e 3º trimestres.

⁸ SAS Forecasting é uma plataforma criada pela empresa SAS para gerar análises de séries temporais e modelos preditivos.

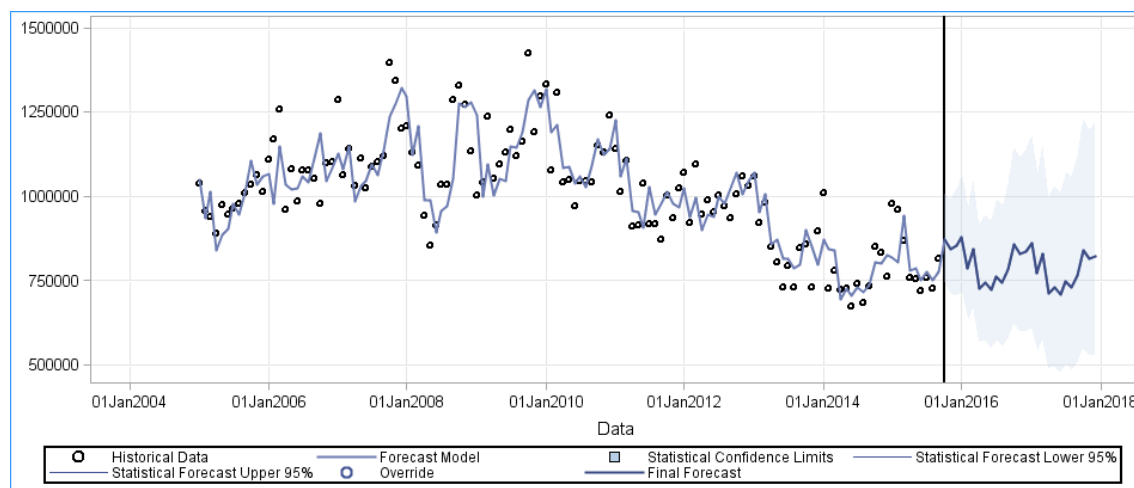


Figura 10 – Predição para o volume de ligações recebidas entre 2016 e 2018

Para a validação desse modelo, verifica-se que essa estimativa apresenta porcentagem do erro médio absoluto (MAPE) muito baixa (5,87%). Com isso, conclui-se que essa função de predição consegue explicar os dados de ligações recebidas pelo *call center* Cemig D com precisão satisfatória. Assim, obtêm-se as predições para o volume de chamadas para 2016, conforme Tabela 9.

Tabela 9 - Predição de chamadas recebidas CEMIG 2016

Mês/Ano	Valor Predito	Limite Inferior (IC 95%)	Limite Superior (IC 95%)	Erro Padrão Predito
jan/16	878.206	716.717	1.065.145	88.987
fev/16	787.675	633.452	968.047	85.469
mar/16	843.795	669.241	1.049.929	97.258
abr/16	726.322	568.535	914.367	88.368
mai/16	742.719	574.105	945.420	94.896
jun/16	722.371	551.682	929.279	96.518
jul/16	762.902	575.910	991.391	106.220
ago/16	744.383	555.662	976.763	107.675
set/16	781.502	577.067	1.035.105	117.140
out/16	855.570	625.120	1.143.509	132.597
nov/16	828.904	599.456	1.117.595	132.556
dez/16	837.036	599.316	1.138.165	137.879

A partir da predição apresentada na Tabela 9, calcula-se a estimativa de ligações recebidas a cada mês, dia e intervalo. O procedimento executado é captar a proporção de chamadas recebidas em cada período em relação ao total de chamadas que entrarão no *call center* Cemig D pelo modelo apresentado na Figura 10.

Por exemplo, no mês de janeiro de 2015, tem-se a proporção relativa de chamadas recebidas para cada dia e intervalo de horário de 30 minutos em relação ao total de ligações

do mês. No de janeiro/2015 foram recebidas 976.894 chamadas. Assim, observa-se que no horário de 09h30min de uma segunda-feira, recebeu-se em média 1.468 ligações, o que representa em média 0,15% do total de ligações.

Esse processo é replicado para cada mês de planejamento de ligações. Com isso, tem-se a representação percentual que cada dia e intervalo de horário de 30 minutos de um ano (como 2015) e replicado para o próximo exercício de planejamento (2016). Portanto, entende-se que a proporção de cada chamada por dia, intervalo de horário de 30 minutos e mês de um ano será a mesma do próximo. Como foi feita a predição de 878.206 ligações para o mês de janeiro/2016, a taxa de ligações recebidas para o intervalo de 09h30min será de $0,15 \times 878.206 = 1.320$ chamadas. Então, essa estimativa será utilizada para o dimensionamento da central telefônica via teoria de filas.

Além da estimativa da taxa de ligações recebidas, calcula-se também o TMA médio para os dias e intervalos de cada mês de apuração. Nesse caso, o TMA médio de 2015 é utilizado como parâmetro para 2016 e utilizado para o cálculo da taxa de serviços realizados. Por exemplo, tomando novamente como base o intervalo de 09h30min de uma segunda-feira do mês de janeiro, tem-se que em 2015 o tempo médio de atendimento foi de 297 segundos, parâmetro que é utilizado para o mesmo período em 2016. Essa prática também é replicada para cada mês de planejamento.

Por fim, admite-se que a taxa de entrada de ligações e a taxa de serviço de 2016 são explicadas pelas estimativas apresentadas até aqui na descrição do algoritmo 2. Esses valores são utilizados como parâmetros da fila M/M/s e posteriormente aplicados ao algoritmo 1. Por exemplo, ao rodar a formulação do programa *Turbo Table*, para o intervalo apresentado para estimar a taxa de chamadas recebidas (1.320 ligações) e o tempo médio de atendimento (297 segundos) para 2016, obtém-se que são necessários 237 pontos de atendimento para manter o INS em 85%.

5.2 Modelo Proposto

Como foco principal deste trabalho, foi construído um programa utilizando a plataforma *R* com o objetivo de executar o planejamento de equipes do *call center* Cemig D com base nos conceitos de filas G/G/s. Com isso, tem-se que a distribuição dos parâmetros que compõem esse tipo de fila pode ser modelada por qualquer distribuição de probabilidade.

Além disso, utilizaram-se as medidas de *performance* de uma fila com chamadas abandonadas com o objetivo de complementar a adequação do regulatório do sistema de dimensionamento de equipes do *call center*. Assim, as estimativas de taxas de abandono são utilizadas em conjunto como o nível de serviço proposto para planejar o número de pontos de atendimentos aderentes às regras do setor elétrico. Para tanto, é necessário ajustar o planejamento de equipes de acordo com o índice mínimo de abandono (4%) e o nível de serviço mínimo (85%) regulados pela ANEEL.

Portanto, o planejamento de equipes da central de teleatendimento Cemig D foi dividido em duas formulações:

1. Algoritmo 3 para o número mínimo de atendentes para um $INS \geq 85\%$;
2. Algoritmo 4 para ponderação do número de atendentes pelo $IAb \leq 4\%$.

Algoritmo 3– Número mínimo de atendentes para um $INS \geq 85\%$

-
1. Informar o mês para o dimensionamento das equipes do *call center*
 2. Definir o nível de serviço $INS \geq 85\%$
 3. **Para** os dias de semana (z) de domingo a sábado **faça**
 4. **Para** os intervalos de horário (i) de 00:00 à 23:30 **faça**
 5. Separar os dados relativos às entradas das ligações no *call center* para o mês informado, dia da semana e intervalo de horário da iteração
 6. Separar os dados relativos ao TMA do *call center* para o mês, dia da semana e horário da iteração
 7. Ajustar os modelos de probabilidade para os dados de entrada de ligações (função *fitdist*)
 8. Executar o teste χ^2 para o modelo de entrada de ligações (função *chisq.test*)
 9. Ajustar os modelos de probabilidade para os dados do TMA
 10. Executar o teste χ^2 para o modelo para o TMA (função *chisq.test*)
 11. Escolher a melhor distribuição para o TMA e entrada de Ligações pelo critério de informação
 12. Ajustar as taxas de entrada (λ) e TMA (μ) de acordo com o melhor modelo de distribuição escolhido no passo 11
 13. Calcular o número inicial de atendentes (S)
 14. Calcular as medidas C_s^2 e C_a^2
 15. Ajustar os valores para as medidas de performance W_q e L_q
-

-
16. **Enquanto** $L_q > \lambda * INS$ **faça**
 17. $S=S+1$
 18. Calcular o novo W_q
 19. Calcular o novo L_q
 20. **Fim enquanto**
 21. Preencher a matriz Agentes $[z,i]$ com o número de atendentes necessários para cada dia e intervalo
 22. **Fim para**
 23. **Fim para**
 24. Exibir a matriz de dimensionamento com base no INS
-

Com esse algoritmo, inicialmente seleciona-se o mês para o qual se deseja executar o planejamento do *call center* (passo 1). Em seguida, é informado o nível de serviço desejado. Como os serviços do *call center* são regulados, o nível de serviço desejado deve ser maior ou igual a 85%.

Nos passos 3 e 4, foi feita uma estrutura de repetição para os dias da semana e para os intervalos de horário. Com isso, o dimensionamento dos atendentes é feito para uma semana de sete dias de trabalho, uma vez que o *call center* da Cemig D deve funcionar durante todos os dias de uma semana e 24 horas por dia. Assim, o resultado é uma matriz 7 x 48, ou seja, 7 dias da semana e 48 intervalos de 30 minutos cada.

Nos passos 5 e 6 são separados o volume de entradas de ligações no *call center* e o tempo médio de atendimento (TMA) do banco de dados de registro de chamadas. O banco de dados utilizado nesta pesquisa possui as informações da central telefônica entre os anos 2012 e 2016⁹ e é organizado por dia da semana, e cada dia é dividido em intervalos de 30 minutos.

Os passos 7 e 8 dizem respeito aos modelos de probabilidade que podem explicar a taxas de entrada de ligações no *call center*, as quais foram ajustadas de acordo com as distribuições Poisson, Normal, Gama e Binomial Negativa pertencentes à família Exponencial. Além desses modelos, foi ajustada a distribuição Exponencial para fins de comparação com as outras funções. Com isso, verificou-se, para cada modelo, o teste χ^2 , com o objetivo de avaliar a aderência dos dados às funções de distribuição escolhidas.

⁹ Dados relativos ao período compreendido entre 01-jan-2016 e 11-ago-2016.

Nesse caso, utilizou-se o nível de significância de $\alpha = 5\%$ e, se o teste retornar um p -valor menor que essa significância, então o modelo de probabilidade será descartado, pois não é capaz de explicar as taxas de entrada de ligações.

Assim como nos passos 7 e 8, foram programados o ajuste dos dados do TMA nos passos 9 e 10 do algoritmo 2. Contudo, utilizaram-se, para modelar os dados da taxa de atendimento pelo TMA, as funções de probabilidade Weibull, Exponencial, Gama, Log-Normal e Normal. Exceto pela Normal, as outras funções são conhecidas por explicarem dados relativos a tempo de vida. Como o TMA pode ser entendido como uma variável que explica o processo de nascimento e morte de um cliente na fila, adotaram-se essas funções por terem uma relação bem próxima com esse tipo de grandeza.

Os passos 7, 8, 9 e 10 foram programados na plataforma R com o auxílio do pacote *fitdistrplus*¹⁰. Este pacote já apresenta funções pré-programadas para o delineamento dos modelos de probabilidades escolhidos. Além disso, é permitido ao pesquisador escolher o método adequado para a estimação dos parâmetros de uma distribuição. Neste trabalho, foram utilizados os métodos dos momentos e da máxima verossimilhança, de acordo com a conveniência de ajuste dos dados trabalhados do *call center* Cemig D.

Caso mais de um modelo ajuste adequadamente as entradas de ligações e a taxa de serviço pelo TMA, será necessário utilizar o método “critério de informação” para seleção do melhor modelo de probabilidade. Pelo critério AIC, decide-se pelo passo 11 que o melhor ajuste para os parâmetros da fila é aquele que apresentar menor valor de informação Akaike.

Por exemplo, na aplicação dessas etapas do algoritmo 3, foi ajustado o número de atendentes para todas as segundas-feiras do mês de janeiro no intervalo de 11h30min às 12h00min. Nesse intervalo, obtém-se como resultado que a distribuição Gama com parâmetros de forma igual a 18,04 e escala igual a 0,011 é a melhor distribuição para explicar a entrada de ligações. Apesar disso, todas as distribuições estudadas para esse intervalo eram válidas para explicar o volume de entrada de ligações, pois os p -valores (acima de 0,2) do teste qui-quadrado foram maiores que o nível de significância proposto. Isso pode ser observado pela figura 11, em que são aplicadas as distribuições de probabilidade e pode-se observar a aderência dos dados amostrados.

¹⁰ Pacote desenvolvido na plataforma R utilizado para modelagem de distribuições.

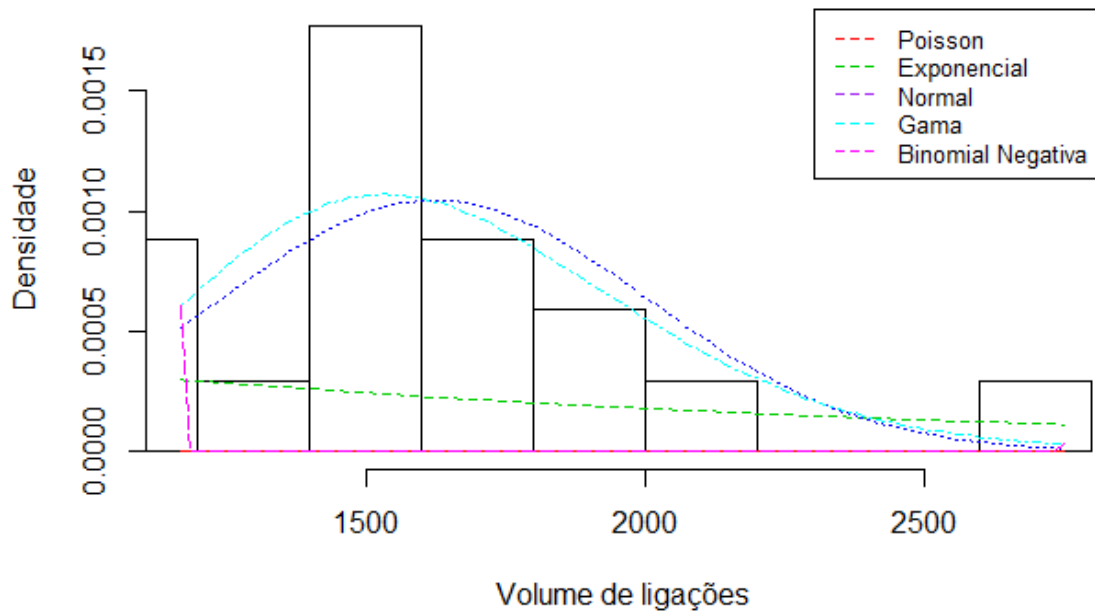


Figura 11 – Modelos de distribuição aplicados ao volume de chamadas atendidas entre 11h30min e 12h00min para as segundas-feiras de janeiro

Além dessa análise, para selecionar o melhor modelo que explique a variável de entrada, utilizou-se o teste de Akaike e o resultado apresentando pelos modelos para realizar essa avaliação. No intervalo 11h30min às 12h00min, a distribuição Gama foi a que obteve menor valor para o teste AIC, cujo resultado foi 251,72.

Da mesma forma, foi analisada a taxa de serviços realizados para esse mesmo período de dimensionamento, que pode ser explicada de acordo com uma distribuição Log-Normal, cujos parâmetros são $\mu_{log} = 1,76$ e $\sigma_{log} = 0,33$. Além disso, o teste AIC retornou como valor o índice 66,28, o menor entre as distribuições investigadas. Todas as distribuições ajustadas podem ser observadas na figura 12.

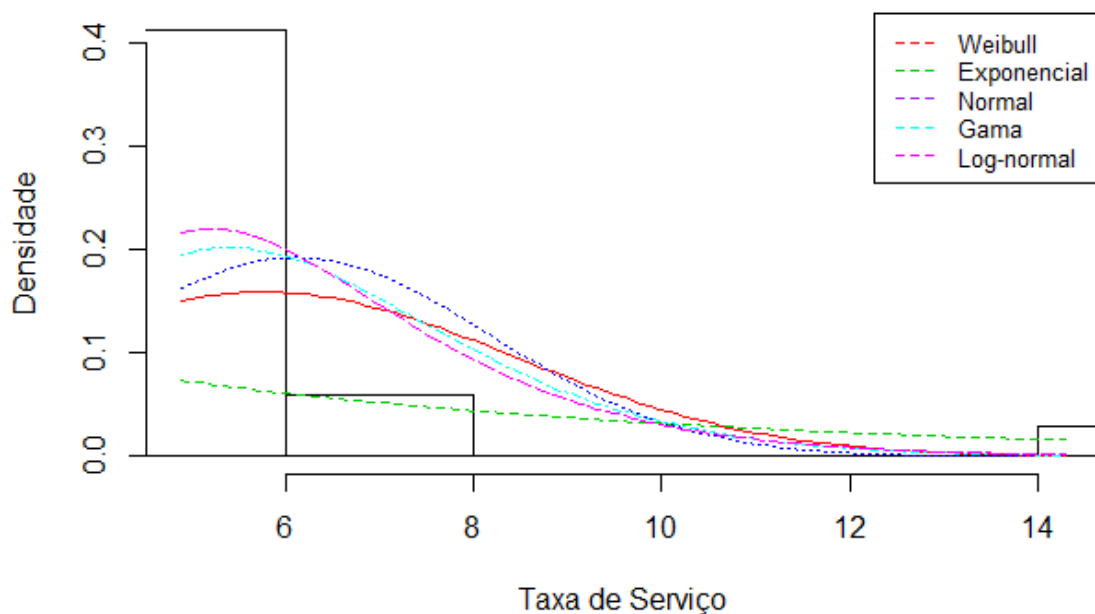


Figura 12 - Modelos de distribuição aplicados ao Tempo Médio de Atendimento entre 11h30min e 12h00min para as segundas-feiras de janeiro

A partir dessas definições, define-se como a taxa de chamadas recebidas e média da distribuição apresentada pela distribuição Gama para o intervalo analisado no exemplo. Já a taxa de serviço é estimada também pela média apresentada pela distribuição Log-normal. Assim, as estimativas utilizadas como parâmetros da fila estão apresentadas na tabela 10.

Tabela 10 – Estimativas para a Taxa de entrada de ligações e Taxa de Serviço – Segunda-feira – entre 11h30min e 12h00min - Janeiro

Parâmetro	Distribuição	Média	Variância
Taxa de entrada de ligações	Gama	1.623	145.981
Taxa de Serviço	Log-normal	6,11	37,49

Após a definição dos modelos a partir dos ajustes apresentados, dos testes de aderência e do critério de informação, calculam-se as taxas de entrada de ligações e a taxa de serviço. Essas medidas são desenvolvidas no passo 12 e representam a média de cada distribuição dos parâmetros da fila G/G/s.

Dadas as taxas calculadas no passo 12, ajusta-se o valor inicial ou valor de teste para o número de atendentes do *call center*. De forma conservativa, calcula-se o menor

número de atendentes em que a intensidade de tráfego seja menor do que 1. Então, seja S o número inicial de atendentes. Assim,

$$S = \left\lceil \frac{\text{Taxa de entrada de ligações}}{\text{Taxa média de atendimento}} + 0,5 \right\rceil \quad (92)$$

e

$$(S) \quad (93)$$

é a quantidade de atendentes arredondada para cima.

Pelo exemplo apresentado, o valor inicial para o número de atendentes necessários para que o sistema de filas seja harmônico, com base no fator de utilização, é dado por

$$S = \left\lceil \frac{1623}{6,11} + 0,5 \right\rceil \approx 266 \quad (94)$$

Com o número inicial de atendentes necessários (S) para que o sistema esteja em equilíbrio, os passos 14 e 15 do programa de planejamento são dedicados ao ajuste das medidas de performance da fila G/G/s. Com isso, calcula-se o tempo médio de espera na fila W_q e o volume esperado de clientes na fila L_q , de acordo com o limite do INS.

Como L_q é dependente de W_q , então foi gerada uma estrutura de repetição que é interrompida somente quando o número de atendentes proporciona valores de performance adequados aos limites regulatórios. Entende-se por quantidade adequada de atendentes o número mínimo necessário de servidores para que o nível de serviço regulatório seja alcançado. Após esse processo iterativo, o valor ajustado para o número de atendentes no *call center* Cemig D para uma segunda-feira, entre 11h30min às 12h00min, do mês de janeiro é igual a 267.

No entanto, o algoritmo 3 só pode garantir que o INS seja alcançado com sucesso, pois somente esse indicador de qualidade do atendimento telefônico é informado até essa fase de execução do programa. Então, incrementou-se a essa formulação de dimensionamento o ajuste do índice de chamadas abandonadas (IAb) para que o número final de atendentes também contemple esse parâmetro. Pela REN ANEEL nº 414/2010, o índice mínimo de chamadas abandonadas não pode ser maior que 4% para um tempo de espera na fila acima de 30 segundos.

Portanto, foi desenvolvido o algoritmo 4 em complementação à formulação que atende somente o nível de serviço.

Algoritmo 4 – Ponderação pelas chamadas abandonadas

-
1. **Para** os dias de semana (z) de domingo a sábado **faça**
 2. **Para** os intervalos de horário (i) de 00:00 à 23:30 **faça**
 3. Separar os dados relativos às chamadas abandonadas no *call center* para o mês informado, dia da semana e intervalo de horário da iteração
 4. Ajustar os modelos de probabilidade para os dados de chamadas abandonadas (função *fitdist*)
 5. Ajustar a probabilidade de abandono de chamada dado um tempo t de espera $P(\text{Abandono}|\text{Espera} > t)$
 6. **Enquanto** $P(\text{Abandono}|\text{Espera} > t) > 4\%$ **faça**
 7. $S = S+1$
 8. Calcular a nova $P(\text{Abandono}|\text{Espera} > t)$
 9. **Fim enquanto**
 10. **Fim para**
 11. **Fim para**
 12. Exibir a matriz de dimensionamento com o número de atendentes ponderados pelo IAb
-

Entre os passos 1 e 3 do algoritmo 4 têm-se formulações similares aos passos 3 a 6 do algoritmo 3. Na ponderação pelo número de chamadas abandonadas, há também a estrutura de repetição limitada pela quantidade de dias da semana e intervalos de horário de um dia de trabalho. Para cada combinação de intervalo e dia da semana, separam-se os volumes de chamadas abandonadas por período.

Após a captação dos dados de abandono, ajusta-se no passo 4 os mesmos modelos de probabilidade utilizados no passo 7 do algoritmo 3. A diferença neste algoritmo 4 é que alguns intervalos e dias possuem períodos em que o volume de abandonos é extremamente alto e pode interferir na execução da modelagem e superestimar a taxa de chamadas abandonadas. Além disso, outros períodos apresentaram volume 0 de chamadas abandonadas e, exceto pela Normal e Poisson, as distribuições escolhidas para modelar o abandono exigem que esta variável seja maior 0.

Para sanar esses problemas, tratou-se primeiramente o banco de dados de abandono de forma a retirar todos os resultados atípicos da amostra. Assim, todo resultado igual ou maior que o primeiro valor atípico é excluído da amostra. Em seguida, foi gerada uma

estrutura de simulação para criar amostras aleatórias de abandono com base em uma distribuição Triangular. Dados os possíveis resultados da amostra ajustada, em que o mínimo valor é 1 por convenção para eliminar os resultados 0, o algoritmo 4 evolui para a fase de modelagem apontada no passo 4 com a amostra produzida.

Ressalta-se também que o teste de aderência dos modelos e a escolha da melhor distribuição para ajustar a taxa de abandono foram utilizados da mesma forma como descrito nos passos 8 e 11 do algoritmo 3. Assim, a taxa de chamadas abandonadas é medida pela esperança pertencente à distribuição que melhor explique esses dados.

No mesmo exemplo do intervalo entre 11h30min e 12h00min, de segunda-feira do mês de janeiro, foram executadas as modelagens de acordo com as distribuições apresentadas na figura 13. Nesse período, a distribuição que melhor explica a taxa de abandonos é o modelo Exponencial, cujo parâmetro de taxa da distribuição é 0,0375.

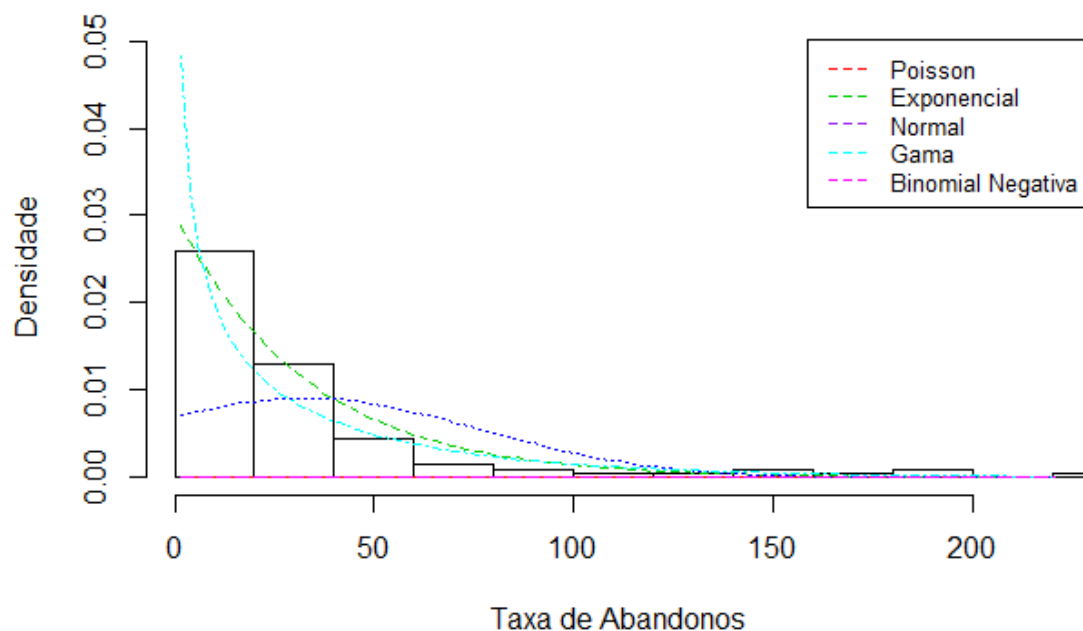


Figura 13 – Taxa de abandonos entre 11h30min e 12h00min para as segundas-feiras de janeiro

Por fim, executa-se nos passos 6 e 8 uma adaptação da função de probabilidade de chamadas abandonadas em relação ao tempo de espera na fila. Essa função foi adaptada do trabalho produzido por Garnett, Mandelbaum e Reiman (2002) em que os mesmos consideram uma fila de atendimento com o elemento abandono.

Pela lógica do algoritmo 4, caso o número de atendentes estimados pelo algoritmo 3 sejam suficientes para não violar o IAb regulatório, então a iteração é interrompida e o

número de atendentes se matém conforme o algoritmo 3. Caso contrário, a iteração é rodada até que o número de atendentes seja suficiente para que o IAb fique abaixo de 4%. No exemplo apresentado, a quantidade de atendentes que atenderá tanto ao INS e ao IAb escolhidos são os mesmos 267.

Assim, para um número S de atendentes, se o volume de chamadas abandonadas for abaixo de 4%, então S é a quantidade ideal que permite ao *call center* alcançar o INS e o IAb propostos pela regulação. Caso contrário, acrescenta-se uma unidade ao número de atendentes até que a função descrita nos passos 6 a 8 seja menor que o índice de chamadas abandonadas exigidas.

5.3 Considerações Finais

Uma das diferenças entre o programa de planejamento G/G/s e a solução do *Turbo Table* apresentado no algoritmo 1 é a forma como são modeladas as entradas de ligações e o TMA. Enquanto o TT utiliza conceitos de filas M/M/s e distribuição dos parâmetros com distribuição Exponencial, no planejamento G/G/s isso não ocorre necessariamente. Portanto, com os ajustes dos parâmetros de acordo com os conceitos de filas generalizadas, pode-se obter estimativas para tempo médio de atendimento e da taxa de entradas mais precisas, pois seus resultados são ajustados de acordo com a característica em termos de distribuição de probabilidades que lhes é peculiar.

Outro ponto de observação é que o *Turbo Table* não faz a ponderação do nº de atendentes em relação ao nº de chamadas abandonadas. Assim, há a possibilidade de que em certos intervalos de tempo a estrutura do *call center* esteja subdimensionada, o que pode acarretar prejuízos regulatórios, pois não há a garantia de que os parâmetros de qualidade estabelecidos regulatórios sejam alcançados.

Portanto, esses dois métodos distintos de dimensionamento tendem a causar efeitos quantitativos também diferentes. Assim, em termos pecuniários, o planejamento de equipes das formulações propostas nos algoritmos 1 e 2 e nos algoritmos 3 e 4 deve ser comparado para definição da estrutura computacional mais eficaz para empresas do setor elétrico brasileiro.

6 COMPARAÇÃO ENTRE OS MÉTODOS DE DIMENSIONAMENTO

Nesta pesquisa, foram desenvolvidos os conceitos e os métodos para o dimensionamento das equipes do *call center* Cemig D baseando-se em teoria de filas. No caso do procedimento atualmente utilizado pelo planejamento da central telefônica, a ferramenta de *Microsoft Excel*, *Turbo Table*, fornece a aplicação de teoria de filas M/M/s para a estimação da quantidade de servidores necessária para o atendimento diário de ligações de tal forma que os limites regulatórios sejam atendidos.

Na proposta alternativa, são trabalhados os conceitos de teoria de filas generalizadas e a estimação dos parâmetros dessa fila baseada em modelagem de funções de probabilidades específicas. Além disso, utilizaram-se outras análises em estatística, como critério de informação e aderência de modelos, para selecionar o melhor modelo que explique os valores trabalhados nessa teoria.

No entanto, é necessário realizar uma análise criteriosa entre os dois métodos propostos e avaliar qual é mais eficaz para o dimensionamento da central telefônica. Para isso, foram implementadas comparações entre os dois métodos a partir dos volumes de chamadas do primeiro semestre de 2016. Assim, é possível apurar a economia realizada ao aplicar a função de dimensionamento baseada em filas generalizadas.

6.1 Algoritmo para definição do dimensionamento

O planejamento das equipes do *call center* Cemig D é feito para atender os seguintes quesitos:

- Estruturar uma régua de atendentes linear para os dias úteis, ou seja, a quantidade de pontos de atendimento de segunda a sexta será a mesma para cada intervalo;
- Dimensionar os sábados e os domingos com o menor número de atendentes possível devido à queda de ligações.

Com isso, foi desenvolvido o algoritmo para definição da régua de atendentes para os dias úteis. Como o número de servidores de segunda à sexta, para os 48 intervalos, deve ser linear, então se optou por avaliar a média e o número máximo de atendentes por horário para definir qual dessas duas métricas é a mais econômica. Como o contrato de prestação de serviços da AeC define que o preço por PA/hora é de aproximadamente R\$ 23,19 (vinte

e três reais e dezenove centavos) ou R\$ 11,60 (onze reais e sessenta centavos) a cada meia hora, aplicou-se a seguinte lógica pelo algoritmo 5:

Algoritmo 5 – Dimensionamento para os dias úteis

-
1. Executar o algoritmo 1 - Dimensionamento Turbo Table
 2. Executar o algoritmo 2 - Número mínimo de atendentes para um $INS \geq 85\%$ + algoritmo 3 - Ponderação pelas chamadas abandonadas
 3. **Para** os intervalos de horário (i) de 00:00 à 23:30 **faça**
 4. Dimensionar o *call center* em dias úteis de acordo com o passo 2 pela média dos intervalos (A)
 5. Dimensionar o *call center* em dias úteis de acordo com o passo 1 pela média dos intervalos (B)
 6. Calcular a diferença em R\$ entre os passos 4 e 5, ou seja, $Dif_média = soma(B * R\$11,60) - soma(A * R\$ 11,60)$
 7. **Fim para**
 8. **Para** os intervalos de horário (i) de 00:00 às 23:30 **faça**
 9. Dimensionar o *call center* em dias úteis, de acordo com o passo 2, pela quantidade máxima de atendentes por intervalos (C)
 10. Dimensionar o *call center* em dias úteis, de acordo com o passo 1, pela quantidade máxima de atendentes por intervalos (D)
 11. Calcular a diferença em R\$ entre os passos 9 e 10, ou seja, $Dif_máximo = soma(D * R\$11,60) - soma(C * R\$ 11,60)$
 12. **Fim para**
 13. **Se** $Dif_média > Dif_máximo$ **então**
 14. Definir A como sendo o dimensionamento ideal
 15. **Senão** Definir C como sendo o dimensionamento ideal
 16. **Fim se**
-

No algoritmo 5, aplicam-se os métodos de dimensionamento elaborados pelos algoritmos 1 – *Turbo Table* e 2+3 – Filas generalizadas + ponderação pelas chamadas abandonadas. Como os dias úteis devem ter a mesma quantidade de atendentes, foram definidos pelos passos 3 a 7 a lógica de dimensionamento por intervalo de horário a partir das médias de atendentes necessários a cada 30 minutos. Além da média, foi calculado

também o dimensionamento a partir dos valores máximos de atendentes por intervalo de horário, conforme definição apresentada entre os passos 8 e 12.

Para avaliação do melhor método, foi estabelecida a regra da diferença entre os valores pecuniários para definição do melhor dimensionamento. Então, o planejamento que apresentar maior diferença será adotado como o mais eficaz. Essa análise está apresentada nos passos 6 e 11 do algoritmo 5. Com essas diferenças, pode-se escolher o modelo de dimensionamento mais econômico. Entre os passos 13 e 15, avaliam-se as regras de cálculo financeiro e aquele método que apresentar maior diferença de planejamento (método da média e método do máximo) será escolhido como o algoritmo como maior nível de redução de gastos.

Já a definição do volume de atendentes necessário para atender aos sábados e aos domingos é dada pela formulação conjunta dos algoritmos 3 e 4. Assim, nas seções 6.2 e 6.3 serão apresentadas as diferenças quantitativas e monetárias entre as metodologias de dimensionamento tratadas neste trabalho.

6.2 Comparação quantitativa entre os métodos de dimensionamento

Nesta seção, foram realizadas comparações entre as quantidades de atendentes necessárias de cada formulação proposta nesta pesquisa. Além disso, para cada formulação apresentada foi rodada a simulação para cada mês de um ano para dimensionar o *call center* Cemig D em dias úteis, sábado e domingo.

Conforme a Figura 14, foram comparados os números de atendentes necessários pela formulação G/G/s e pela formulação *Turbo Table*. Observou-se que, no período de um ano, o algoritmo que utiliza os conceitos de filas generalizadas demandou a cada intervalo de 30 minutos de um dia útil, em média, a redução em 9% do número de pontos de atendimentos contratados em relação à solução via *Turbo Table*. Por exemplo, no mês de julho, enquanto a solução proposta necessita de pelo menos 4.851 PA em um dia útil para atender aos quesitos regulatórios, a solução via TT demandou 5.425 PA para alcançar as mesmas metas de INS. No total, a formulação G/G/s demanda em um dia útil em um ano 69.776 PA, enquanto a outra formulação requer 76.192 pontos de atendimento para atender aos mesmos critérios.

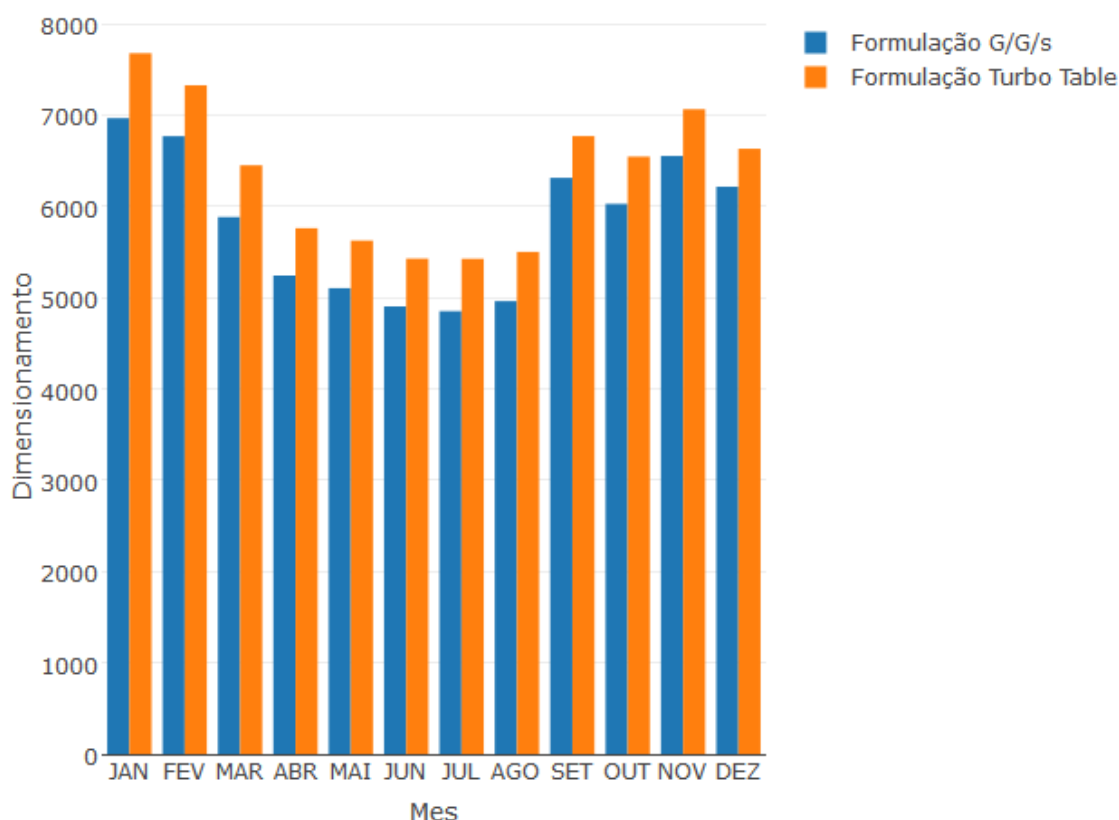


Figura 14 – Comparação entre as metodologias de dimensionamento – Dias úteis

Em um sábado, observou-se que, na formulação via filas generalizadas, há em média uma queda de 6,5% no número necessário de atendentes, ao se comparar com os resultados alcançados pelo *Turbo Table*. Enquanto isso, na soma de um sábado por mês em um ano são necessários 36.376 atendentes no dimensionamento com filas G/G/s, para atender a mesma demanda e nos resultados regulatórios são necessários 38.742 PA pela solução *Turbo Table*. Assim, podem-se verificar esses resultados visualmente pela Figura 14. O único mês em que a formulação via G/G/s gerou superdimensionamento foi em dezembro, em que houve um aumento aproximado de 4% na quantidade de atendentes necessária em relação do TT.

Na Figura 15, tem-se a comparação relativa ao dimensionamento feito para cada domingo. Nesse dia, tem-se a menor diferença entre as formulações estudadas nesta pesquisa. A redução percentual no volume de PA necessárias na formulação G/G/s é menor que 1% (aproximadamente 0,86%), em relação à função de dimensionamento via TT. Observa-se que nos meses pertencentes ao período úmido, a formulação G/G/s apresenta dimensionamento relativamente maior do que a formulação via *Turbo Table*. Por exemplo, em novembro, a diferença a mais na solução via filas generalizadas foi de 163 PA. Já no

período seco do ano, houve em média uma redução de 102 pontos de atendimentos com a solução G/G/s, mas, no mês de setembro, houve superdimensionamento de 96 PA.

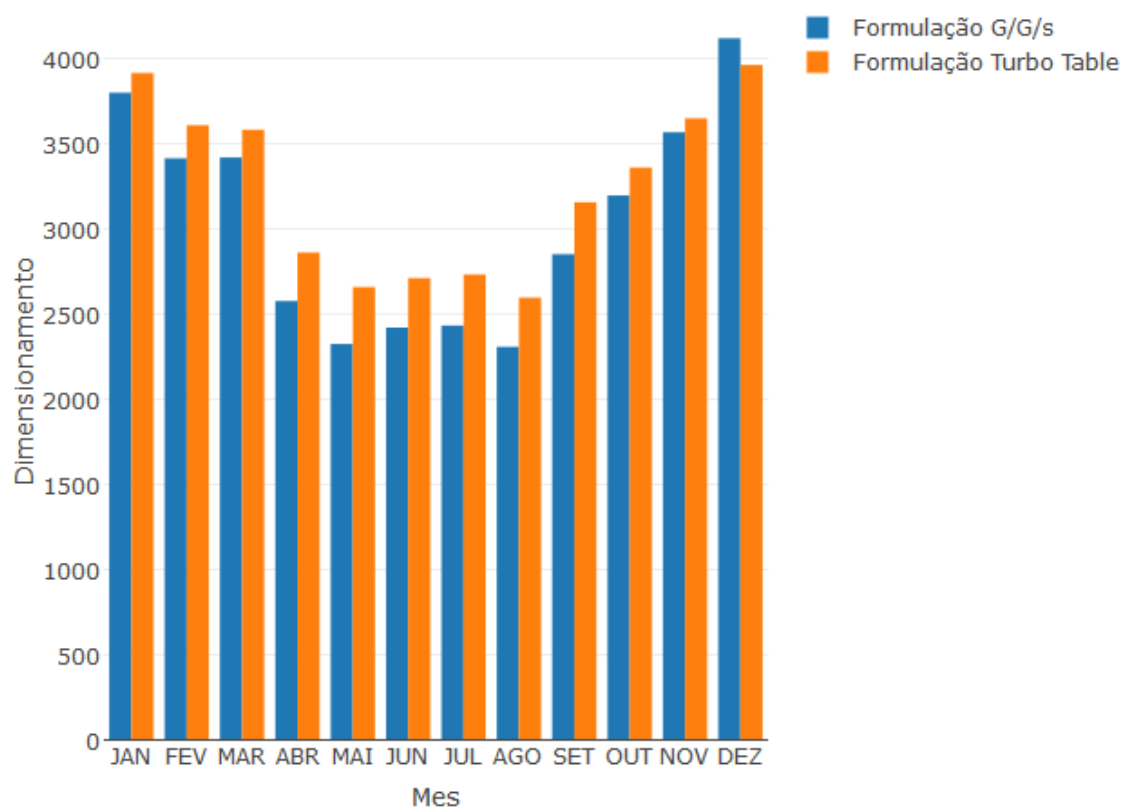


Figura 15 - Comparação entre as metodologias de dimensionamento – Sábado

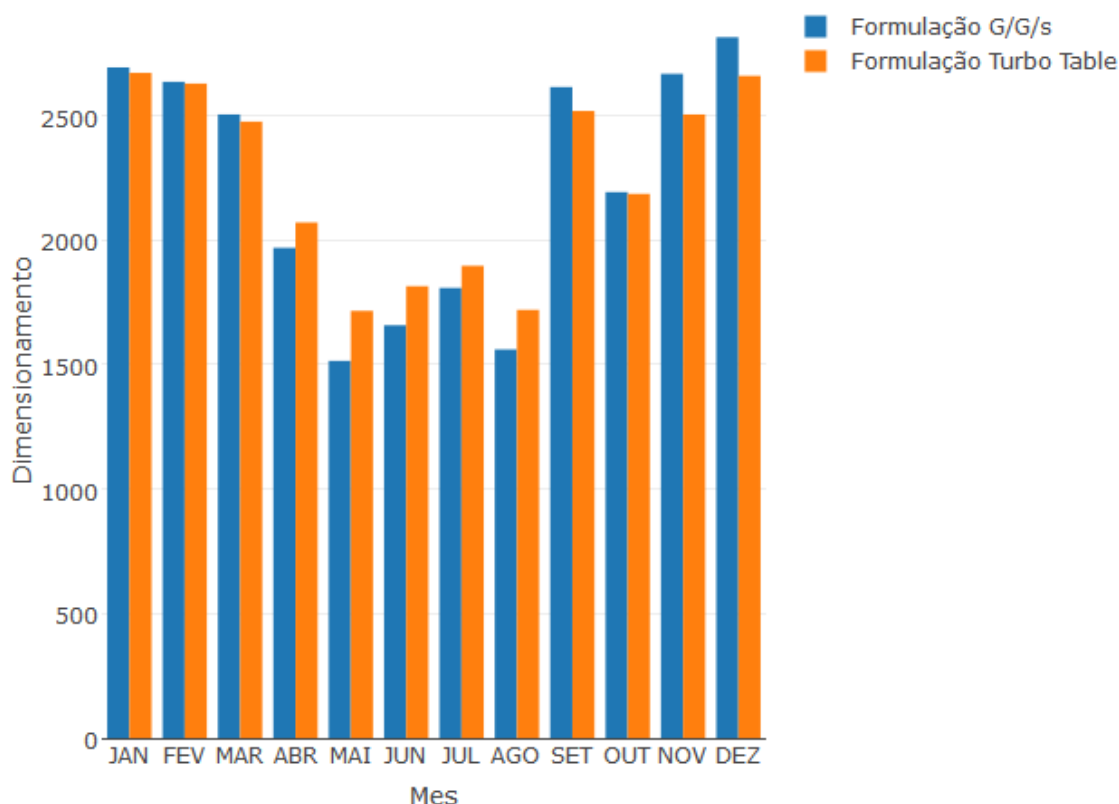


Figura 16 - Comparação entre as metodologias de dimensionamento – Domingo

Além das comparações por dia de trabalho no *call center*, pode-se fazer também o agrupamento de dimensionamento relativo aos períodos do ano (úmido e seco). No período, conjunto de meses relativos aos 1º e 4º trimestres de cada ano, verifica-se que o número de atendentes necessários é 5% maior na solução TT do que na proposição via filas generalizadas da Tabela 11. Nessa época do ano, são necessários mais servidores para captarem as ligações telefônicas, pois há mais chamados registrados devido às ocorrências emergenciais causadas pelo aumento das chuvas nesse período.

Já no período seco, fase compreendida entre o 2º e o 3º trimestre de cada ano, a redução do volume de pontos de atendimento pela metodologia G/G/s, em relação ao *Turbo Table*, foi maior. Entre os meses de abril e setembro, houve uma redução no número de PA de aproximadamente 10%.

Tabela 11 – Dimensionamento semanal por período do ano

Período	Dimensionamento Filas G/G/s (1)			Dimensionamento Turbo Table (2)			% □ (1) - (2)
	Dias úteis	Sábado	Domingo	Dias úteis	Sábado	Domingo	
Úmido	38.406	21.490	15.511	41.687	22.052	15.128	5%
Seco	31.370	14.886	11.126	34.505	16.690	11.738	10%

Por fim, pelos resultados apresentados até então, foram representadas na Tabela 12 as diferenças entre os dimensionamentos por mês e dias úteis pelas duas formulações. Os valores positivos indicam que a solução via *Turbo Table* apresenta planejamento para os pontos de atendimento superestimados em relação à formulação via filas generalizadas. Por exemplo, a diferença em fevereiro é 557, ou seja, a solução via TT é 557 unidades maior do que o planejamento feito pela fila G/G/s.

Além desses resultados, tem-se que as maiores diferenças estão apresentadas no mês de janeiro, nos dias úteis, em que podem ser verificadas, na Figura 17, as diferenças de dimensionamento por intervalos de 30 minutos. Para o sábado e o domingo, estão representados, nas Figuras 13 e 14, respectivamente, os resultados mais expressivos observados no mês de maio, com redução de 334 PA (sábado) e 201 PA (domingo), ao ser comparada a solução via filas generalizadas com o *Turbo Table*.

Tabela 12 – Diferenças entre os dimensionamentos – por mês

Dia	jan	fev	mar	abr	mai	jun	jul	ago	set	out	nov	dez
Dia Útil	711	557	567	519	521	524	574	539	458	518	512	416
Sábado	116	194	163	285	334	291	300	289	305	164	82	-157
Domingo	-22	-7	-29	102	201	157	89	159	-96	-7	-163	-155

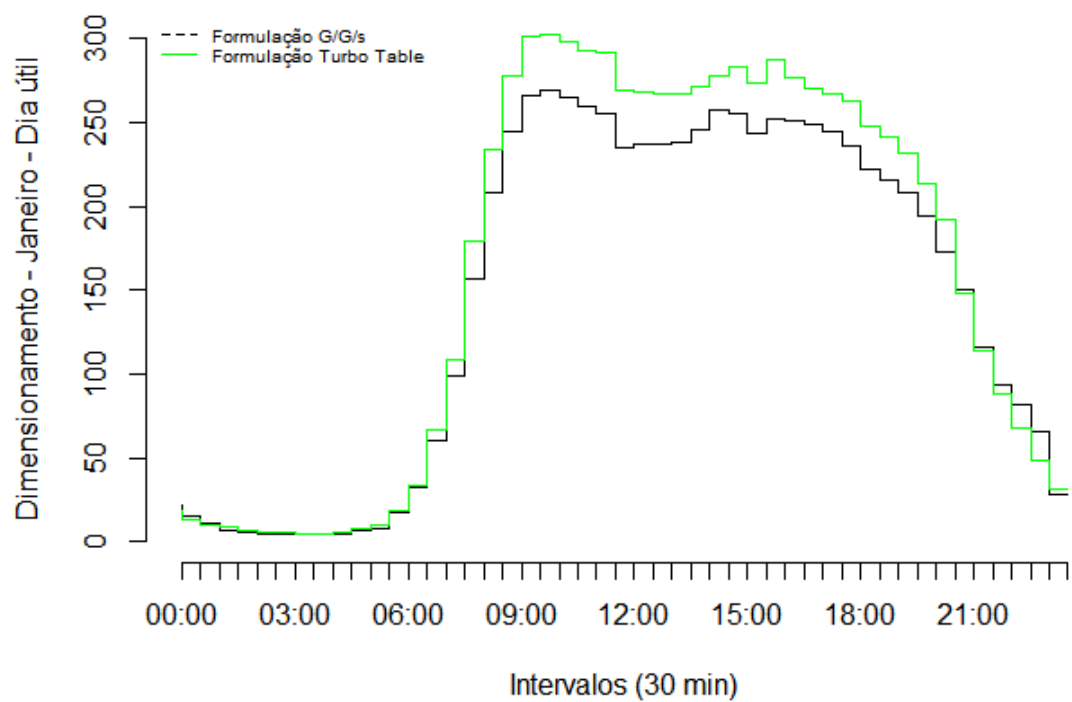


Figura 17 – Dimensionamento para dias úteis – Janeiro

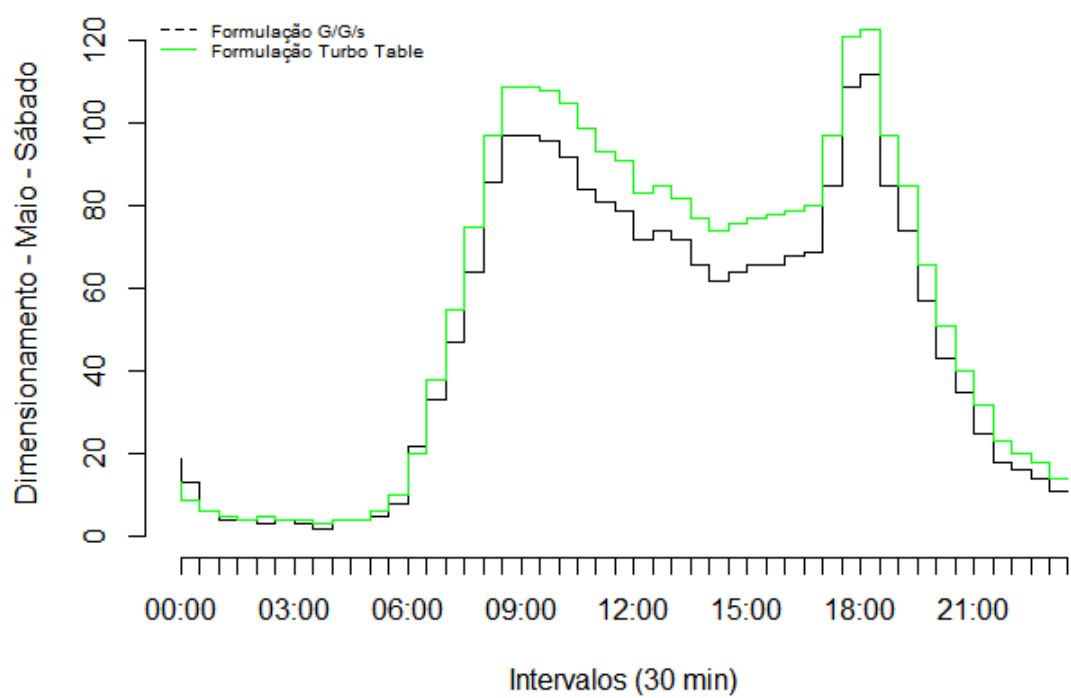


Figura 18 – Dimensionamento para o sábado – Maio

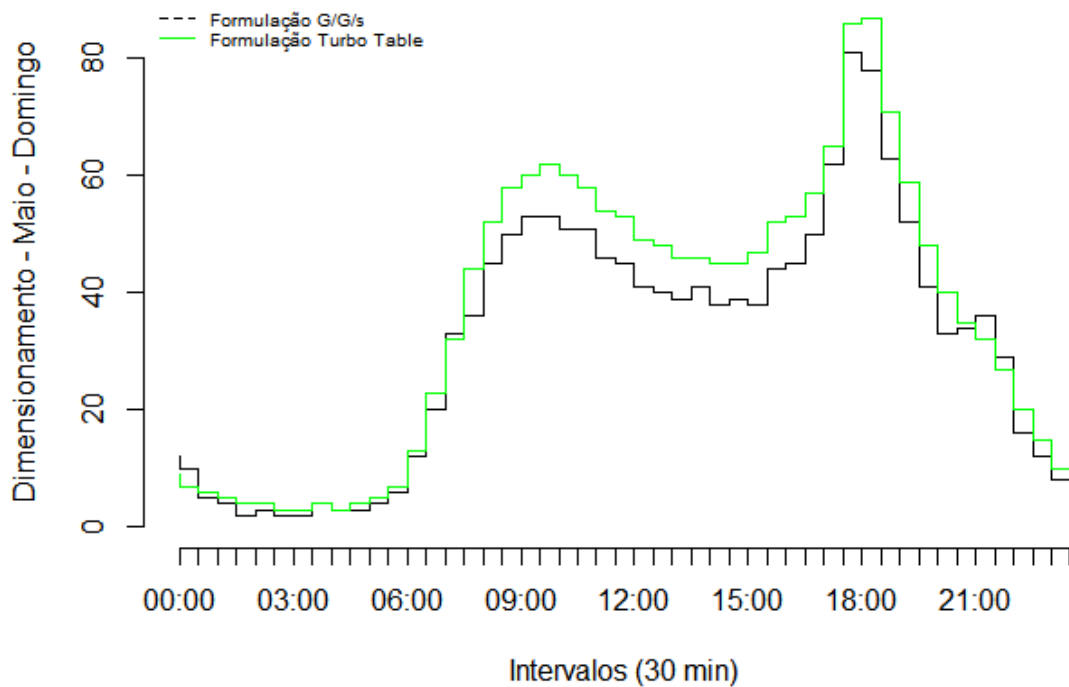


Figura 19 – Dimensionamento para domingo – Maio

6.3 Teste de eficácia do modelo de filas generalizadas: Resultados para o 1º semestre de 2016

Os procedimentos de dimensionamento de equipes do *call center* Cemig D levam em consideração a predição do volume de chamadas para um determinado mês, a representatividade de cada intervalo em relação ao total de ligações previstas e a utilização do algoritmo do *Turbo Table*. Assim, o planejamento da central de atendimento depende do banco de dados histórico da empresa e das proporções de chamadas por cada dia da semana e por período, seja ele úmido ou seco.

Na aplicação do *Turbo Table*, entende-se nesse caso que os parâmetros de entrada de ligações e de tempo de serviço (ou atendimento) são provenientes das estimativas feitas pelo modelo de predição adotado. Com isso, tem-se uma adaptação dos valores calculados ao modelo de fila M/M/s utilizado na ferramenta de dimensionamento, cujos resultados são historicamente eficientes, pois raramente houve períodos em que os indicadores de qualidade da ANEEL não foram alcançados.

Contudo, vislumbra-se a necessidade de comparar o algoritmo criado para o dimensionamento da central de atendimento via filas generalizadas, em relação ao método de dimensionamento atualmente aplicado. Para tanto, todas as estimativas para os

parâmetros das filas G/G/s e M/M/s foram realizadas com base nos dados do *call center* Cemig D até o ano de 2015. Assim, o volume de ligações e os tempos médios de atendimento de 2016 foram utilizados para validação dos procedimentos de dimensionamento, em que se buscou o modelo com menor custo e menor risco de descumprimento das metas regulatórias.

Em relação às regras que a ANEEL determina às distribuidoras de energia, determinou-se como modelo mais eficiente àquele que atende o maior nível de INS possível com menor custo observado. Dessa maneira, implementou-se uma análise de probabilidade em relação aos resultados de 2016 com os dimensionamentos propostos para analisar o nível de serviço alcançado pelos dois métodos. Segundo Gross *et al.* (2008), é possível estimar a probabilidade condicional do tempo de espera em uma fila M/M/s, em que

$$P(T_q > t | T_q > 0) = e^{-(s\mu - \lambda)t} \quad (95)$$

para T_q como o tempo de espera na fila, s é o número de atendentes dimensionados, μ é taxa de serviço para o período e λ é o volume de entradas na fila. Assim, tem-se que essa função afere a probabilidade do tempo de espera ser maior que um valor contínuo, dado que há espera para atendimento dos elementos que estão na fila.

Para aplicação de função de probabilidade, foi estabelecido o dimensionamento para cada dia e intervalo de 30 minutos. Utilizou-se nesse caso a regra de que os números de pontos de atendimento para os dias úteis (de segunda à sexta) seriam os mesmos e que haveria diferença somente nos sábados e nos domingos, pois os volumes de atendimento nos finais de semana são menores. Além disso, optou-se pelo número máximo de atendentes dimensionados para os dias úteis. Ou seja, para cada um dos 48 intervalos, o valor máximo planejado para um determinado dia é aplicado na análise. Por exemplo, considere o intervalo das 09:30 da manhã de um dia útil do mês de janeiro: tem-se que na segunda foram estimados 274 pontos de atendimento, na terça 250, na quarta 260, na quinta 251 e na sexta 237. Para não correr riscos de perda de indicadores regulatórios, para esse intervalo serão planejados 274 atendentes para todos os dias úteis.

Já para as taxas de atendimento e de chegada de ligações no *call center*, foram considerados os resultados reais do 1º semestre de 2016. Por exemplo, no dia 23/03/2016, às 12h00min, foram recebidas 1.009 ligações com tempo médio de atendimento de 294 segundos. Então, a taxa média de ligações recebidas (λ) foi de 1.009 ligações/30 min e a taxa de serviço (μ) 6,12 ligações/30 min. Com o planejamento de 221 PA pelo *Turbo*

Table e 215 PA pelo procedimento via filas generalizadas, tem-se que a probabilidade condicional de atendimento acima de 30 segundos desses é de aproximadamente 0,32% e 0,60%, respectivamente.

O tempo contínuo t foi padronizado para todos os intervalos em 30 segundos, pois se deseja, com essa probabilidade, estimar o volume de chamadas atendidas abaixo desse valor. Assim, tem-se o complementar da probabilidade descrita na fórmula (95) como a proporção de clientes que esperam menos de 30 segundos na fila.

Utilizando o mesmo dia e intervalo (23/03/2016, às 12h00min), tem-se que o número de chamadas atendidas em até 30 segundos é de 1.006 com o dimensionamento do *Turbo Table* e 1.003 com a formulação em filas G/G/s. Nessa análise, ainda é possível estimar o INS alcançado pela central de atendimento. O nível de serviço pode ser calculado com a divisão do número de chamadas atendidas em até 30 segundos pela diferença do total de chamadas recebidas e as abandonadas em até 30 segundos. Nessa análise, tem-se que o total de chamadas recebidas foi de 1.009 ligações e não houve chamadas abandonadas. Portanto o INS para as formulações *Turbo Table* e G/G/s foi de aproximadamente 99%. Com isso, pode-se observar que, com 6 pontos de atendimento a menos na formulação de filas generalizadas, foi possível alcançar os mesmos resultados de nível de serviço que o TT.

Esse mesmo processo foi replicado para os 8.736 intervalos de atendimento entre 01/01/2016 e 30/06/2016 e verificado os resultados regulatórios a partir dos períodos típicos do atendimento telefônico. Pelos resultados apurados na Tabela 13, observa-se que o total de pontos de atendimento no 1º trimestre de 2016, bem como o número de atendentes necessários planejados com as filas generalizadas é similar ao dimensionamento proposto pela formulação *Turbo Table*. Contudo, no 2º trimestre, há uma redução de 6% no procedimento adotado das filas G/G/s. Em termos de custo, verifica-se que no período analisado, houve uma redução de 3% na contração dos pontos de atendimento, como pode ser observado na Tabela 14.

Tabela 13 – Comparação entre os procedimentos para dimensionamento do call center Cemig D

Mês	Chamadas Recebidas (A)	Abandonos abaixo de 30 s (B)	(A) - (B)	Chamadas atendidas em 30 segundos - Estimativa		INS Previsto		Total de PA/30 min	
				MMs TT	GGs	MMs TT	GGs	MMs TT	GGs
Janeiro	761.393	5.043	756.350	610.934	708.393	81%	94%	160.829	171.877
Fevereiro	693.520	5.852	687.668	603.134	612.822	88%	89%	172.072	158.473
Março	729.667	3.456	726.211	640.867	665.256	88%	92%	168.862	169.618
Abril	631.459	2.183	629.276	558.369	565.695	89%	90%	139.741	137.940
Maio	636.134	1.585	634.549	592.285	578.665	93%	91%	161.584	142.954
Junho	641.334	3.049	638.285	588.609	560.657	92%	88%	142.060	133.918
Total	4.093.507	21.168	4.072.339	3.594.197	3.691.489	88%	91%	945.148	914.780

Tabela 14 – Custo total dos pontos de atendimento

Custo de dimensionamento (em R\$) Custo unitário: R\$ 11,60 PA/30 min			
MMs TT		GGs	
R\$	1.865.616,40	R\$	1.993.773,20
R\$	1.996.035,20	R\$	1.838.286,80
R\$	1.958.799,20	R\$	1.967.568,80
R\$	1.620.995,60	R\$	1.600.104,00
R\$	1.874.374,40	R\$	1.658.266,40
R\$	1.647.896,00	R\$	1.553.448,80
R\$	10.963.716,80	R\$	10.611.448,00

Além disso, verificou-se que, pela formulação das filas generalizadas, o nível de serviço acumulado do 1º semestre de 2016 é de 91%, e que em nenhum dos meses houve violação da meta do INS. Nesse mesmo período, o nível de serviço do procedimento de dimensionamento do *Turbo Table* foi de 88%, mas perdendo o resultado do mês de janeiro/2016 (81%).

Em relação ao dimensionamento mensal, tem-se também que há concentração de atendentes por tipo de dia em que é feito o dimensionamento. De acordo com os dados apresentados nas Tabelas 15 e 16, 85% do total de pontos de atendimento se concentram nos dias úteis, na formulação TT, e 84%, na G/G/s. Nesses dias, tem-se também que o INS previsto pelo planejamento se encontra acima de 85%.

Em relação a sábados e domingos/feriados, há uma redução no volume de PA planejadas devido ao baixo volume de chamadas recebidas nesses dias. Observa-se também que o nível de serviço nesses dias no 1º semestre não atingiu o limite proposto pela ANEEL. No entanto, o fato de o planejamento feito para os dias úteis ultrapassar os INS de 85% na maior parte dos meses permite que haja um contrabalanço em relação aos finais de semana. Portanto, não haverá prejuízo regulatório para a Cemig D com os cenários apresentados.

Ainda de acordo com as Tabela 15 e 16, verifica-se que o dimensionamento para o 1º semestre de 2016 em dias é 5% maior na formulação via TT do que na G/G/s. Em compensação, planeja-se um número de PA nos finais de semana na formulação via filas generalizadas aproximadamente 7% maior do que na solução *Turbo Table*.

Tabela 15 – Planejamento de equipes por dia útil e mês

Mês	Dia útil			
	MMs TT		GGs	
	PA/30min	INS	PA/30min	INS
Janeiro	127.138	80%	138.395	95%
Fevereiro	148.328	93%	133.213	93%
Março	147.602	94%	144.783	94%
Abril	118.373	98%	112.243	95%
Maio	142.485	98%	123.716	96%
Junho	125.391	95%	117.153	90%
Total	809.317	93%	769.503	94%

Tabela 16 – Planejamento de equipes por dia de fim de semana e mês

Mês	Sábado				Domingo/Feriado			
	MMs TT		GGs		MMs TT		GGs	
	PA/30min	INS	PA/30min	INS	PA/30min	INS	PA/30min	INS
Janeiro	17.225	80%	17.589	87%	16.466	89%	15.893	87%
Fevereiro	11.284	63%	12.651	81%	12.460	62%	12.609	61%
Março	9906	63%	11.912	86%	11.354	56%	12.923	73%
Abril	9.544	77%	10.996	89%	11.824	37%	14.701	55%
Maio	9.610	85%	9.779	81%	9.489	54%	9.459	51%
Junho	10.124	75%	9.920	71%	6.545	72%	6.845	72%
Total	67.693	74%	72.847	83%	68.138	61%	72.430	66%

6.4 Considerações finais

Pelos resultados apresentados, verifica-se que o procedimento de dimensionamento por meio de filas generalizadas se mostrou mais eficaz para o planejamento das equipes do *call center* Cemig D. Tanto em relação ao *Turbo Table* teórico da seção 6.2 quanto na formulação *Turbo Table* com predição do parâmetro de entrada, os resultados apresentam o custo para planejamento das equipes da central de atendimento telefônico mais elevado do que a solução via G/G/s.

A formulação via filas generalizadas traz uma carga maior de pontos de atendimento nos dias úteis, o que permite suavizar o dimensionamento das equipes nos finais de semana. Com essa abordagem, é possível reduzir a necessidade de atendentes nos períodos em que o atendimento é reduzido significativamente devido à baixa procura por parte dos clientes da distribuição.

Em termos conceituais, foram adaptadas medidas de performance e de probabilidade das filas M/M/s para confirmar os melhores resultados do planejamento das equipes via filas generalizadas. Na literatura consultada, verificam-se as medidas de probabilidades condicionais de tempo de espera como fundamentais para a estimação dos indicadores de qualidade do atendimento telefônico, tais como o INS e IAb.

7 CONCLUSÕES

Diante de um quadro de decréscimo do volume de chamadas observadas no *call center* Cemig D apresentado no capítulo 4 deste trabalho e da necessidade constante de adequação dos processos da distribuição de energia aos ritos regulatórios, surge a oportunidade de desenvolver ferramentas mais precisas para o planejamento de equipes de atendimento telefônico. Nesse caso, é possível elaborar técnicas que rompem os métodos clássicos de predição do volume de chamadas e análises descritivas para estimação do tempo médio de atendimento. Observa-se, além disso, que as ferramentas que utilizam os procedimentos de filas M/M/s condicionam a equipe de planejamento a definir que os parâmetros de análise dessa fila tenham uma característica singular de distribuição, ou seja, só podem ser modelados conforme um modelo Exponencial.

Pelo método atual de dimensionamento do capítulo 5, foi possível observar que atualmente o planejamento das equipes de teleatendimento Cemig D é realizado com base em modelos de predição via alisamento exponencial, o que pode gerar volumes de chamadas superestimadas em relação aos acontecimentos reais. Assim, a estimativa desse parâmetro pode culminar em excesso de pontos de atendimento necessários para captar 85% das chamadas em até 30 segundos. Além disso, o método apresentado pelo *Turbo Table* desconsidera a condição de chamadas abandonadas. Dessa forma, não há perspectiva implementada nesse sistema que assegure que 4% das chamadas recebidas na central abandonem o atendimento.

Já o algoritmo proposto para solucionar o problema de dimensionamento, baseado em filas generalizadas e modelagem estatística, apresentou-se mais eficaz em relação ao algoritmo *Turbo Table*. Verificou-se uma melhor distribuição de pontos de atendimento ao longo das semanas com menor número de atendentes necessários para alcançar os níveis de qualidade de atendimento demandados pelo regulador de energia elétrica no Brasil. Além disso, o fato de estimar os parâmetros de uma fila generalizada, de acordo com a função probabilística mais apropriada, proporcionou menor risco de superestimação do volume de chamadas recebidas e do tempo de serviço.

Cabe lembrar, ainda, que a solução via filas G/G/s ainda é pouco explorada para o contexto analítico desenvolvido nesta pesquisa. Nessa teoria, há poucos conceitos desenvolvidos para construir algumas medidas de *performance* do sistema que é estudado. Por exemplo, funções de probabilidade de tempo de espera na fila são bem desenvolvidas

em modelos clássicos de filas, tais como a M/M/s. Em filas generalizadas, é compreensível a ausência de modelos desse tipo, pois, para cada característica dos dados de entrada e do tempo de serviço, haverá um modelo de distribuição diferente da Exponencial. Portanto, é possível desenvolver uma medida de performance singular para cada parâmetro da fila G/G/s, de acordo com as funções investigadas que melhor expliquem essas medidas.

Além disso, como o programa foi todo desenvolvido em plataforma R, a aplicação de funções existentes para a modelagem dos parâmetros da fila auxiliou na composição lógica da solução via G/G/s. Assim, o desenvolvimento associado ao programa de dimensionamento pode ser transformado em um pacote R e disponibilizado para outros pesquisadores e empresas do ramo de energia utilizarem.

7.1 Sugestões para trabalhos futuros

Este trabalho não pode ser considerado exaustivo, pois há outras possibilidades de desenvolvimento de pesquisa em termos de filas G/G/s. Nesse sentido, pesquisas futuras podem fomentar os conceitos de filas de generalizadas, tais como:

- Desenvolvimento de medidas de *performance*, como probabilidade condicional de tempo de espera e probabilidade condicional de tempo de espera,
- Esperança do número de agentes ocupados e do número de clientes aguardando atendimento;
- Criação de funções de probabilidades específicas para filas generalizadas com abandono (probabilidade de um cliente abandonar a fila, dado que está esperando atendimento, e de abandono, por exemplo) com as respectivas esperanças e variâncias;
- Por fim, o desenvolvimento de uma fila do tipo G/G/ s + M, que significa uma fila cujos parâmetros de entrada (λ) e tempo de serviço (μ) são distribuições generalizadas, capacitadas com s atendentes e que admitem uma taxa θ de abandonos.

8 REFERÊNCIAS BIBLIOGRÁFICAS

AKAIKE, H. A. A new look at the statistical model identification. IEEE Transactions on Automatic Control, Notre Dame, v. 19, n. 6, p. 717-723, 1974.

BOWERMAN, B. L.; O'CONNELL, R. T. Forecasting and time series: an applied approach. California: Duxbury Press, 1993. 848p. p. 379-426.

BRASIL. Agência Nacional de Energia Elétrica – ANEEL. Resolução Normativa N° 414/2010. Disponível em: <<http://www.aneel.gov.br/cedoc/ren2010414.pdf>>. Acesso em: 18 jul. 2015.

BURNHAM, K. P.; ANDERSON, D. R. Multimodel inference: understanding aic and bic in model selection. Sociological Methods and Research, Beverly Hills, v. 33, n. 2, p. 261-304, 2004.

CHATFIELD, C. The analysis of time series: an introduction. Florida: Chapman & Hall, 2003. 357p. p. 76-80.

DEGROOT, M.H.; SCHERVISH, M. J. Probability and Statistics. Massachusetts: Pearson Education, 2011. 911p. p. 93 – 346.

FAHRMEIR, L.; TUTZ, G. Multivariate Statistical Modelling Based on Generalized Linear Models. New York: Springer, 2001. 518p. p. 15-24.

FORBES, C.; EVANS, M.; HASTINGS, N.; PEACOCK, B. Statistical Distributions. New Jersey: John Wiley & Sons, 2011. 231p. p. 62 – 156.

GARNETT, O., MANDELBAUM, A., REIMAN, M. I. Designing a Call Center with impatient customers. In: Manufacturing & Service Operations Management. INFORMS, 2002, p. 208-227.

GAUTAM, N. Analysis of queues: methods and applications. Florida: CRC Press, 2012. 788p. p. 151-200.

GROSS, D.; SHORTLE, J. F.; THOMPSON, J. M.; HARRIS, C. M. Fundamentals of queueing theory. Massachusetts: John Wiley & Sons, 2008. 528p. p. 49-356.

HILLIER, F. S.; LIEBERMAN, G. J. Introduction to Operations Research. New York: McGraw-Hill Education, 2015. 1411p. p. 731-799.

HYNDMAN, R. J. Forecasting with exponential smoothing: the state space approach. Berlin: Springer, 2008. 359p. p. 11-17.

- HOSHMAND, A. R. Business Forecasting: a practical approach. New York: Taylor & Francis, 2010. 382p. p. 9-29.
- LEHMANN, E.L.; CASELLA, G. Theory of Point Estimation. New York: Springer, 1998. 617p. p. 23-44.
- MAKRIDAKIS, S. G.; WHEELWRIGHT, S. C.; HYNDMAN, R. J. Forecasting: methods and applications. New York: John Wiley & Sons, 1997. 656p. p. 29-34.
- MANDELBAUM, A.; ZELTYN, S. Service engineering in action: The Palm/Erlang-A Queue, with applications to call centers draft. Haifa, Israel. Disponível em: <http://sistemas-humano-computacionais.wdfiles.com/local--files/capitulo%3Asistemas-de-servico/mandelbaum07.pdf>. Acesso em 8 fev 2016.
- MONTGOMERY, D.C.; RUNGER, G.C. Applied Statistics and Probability for Engineers. New York: John Wiley & Sons, 2002. 822p. p. 16-140.
- RAYNER, J. C. W.; THAS, O.; BEST, D. J. Smooth Tests of Goodness of Fit. Singapore: John Wiley & Sons, 2009. 290p. p. 17-26.
- TAYLOR, J.R. An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements. California: University Science Books, 1997. 343p. p. 271-273.
- WHITT, W. Sensivity of performance in the Erlang A model to changes in the model parameters. In: European Journal of Operational Research. Elsevier, 2006, p. 247-260.

ANEXO

O programa de dimensionamento de equipes de *call center* foi desenvolvido na plataforma R-Studio (Version 0.99.489). Este programa retorna à matriz de dimensionamento para um determinado mês (*Mo*), ano (*year*), Nível de Serviço (*SL*), Índice de Abandonos (*Ab*) e para um determinado número de simulações de chamadas abandonadas (*simu*). A variável *SL* tem grandeza decimal e está compreendida entre 0 e 1. Já a variável *Ab* está compreendida entre 0 e 0,04.

```
1. #####
2. # Planejamento de equipes de atendimento para o call center CEMIG-D baseado em um algoritmo de filas
   generalizadas#
3. # Mestrado em Engenharia de Produção
4. # Mestrando: Alexandre Henrique Martins
5. #Orientador: Anderson Laécio Galindo Trindade
6. #####
7. rm(list=ls(all=TRUE))
8. #chamando o banco de dados de chamadas#
9. dimensionamento1=function(Mo,year,SL,Ab,simu){
10. #####
11. ## Leitura do bando de dados##
12. #####
13. bd<-read.csv("C:\\Back up_Alexandre\\Mestrado\\Produção\\Dissertação\\Bancos de Dados\\chamadas.csv",
   header = TRUE, sep = ";")
14. #####
15. ## Mês de dimensionamento, INS e IAb##
16. #####
17. Mes<-Mo
18. INS<-SL
19. y<-1-INS
20. IAb<-Ab
21. Pa=0
22. if(Mes==1 || Mes==2 || Mes==3 || Mes==10 || Mes==11 || Mes==12) {Pa=1}
23. #bd#
24. #transformando o bd em matriz#
25. base<-array(bd)
26. require(fitdistrplus)
27. require(gmp)
28. require(triangle)
29. tempo<-matrix(nrow=7,ncol=48)
30. Abandonada<-matrix(nrow=7,ncol=48)
31. Prob_aban<-matrix(nrow=7,ncol=48)
32. Prob_aban_d<-matrix(nrow=7,ncol=48)
33. Agentes<-matrix(nrow=7,ncol=48)
34. #####
35. ## Loop de dimensionamento ##
36. #####
```

```

37.   for (z in 1:7){
38.     for (i in 1:48){
39.       #lógica para filtrar o volume de chamadas recebidas#
40.       Dia_m<-matrix(subset(base[,4],base[,3]==z & base[,16]==Mes & base[,14]!=4 & base[,17]<year),
nrow=48)
41.       m<-Dia_m[i,][Dia_m[i,]>0]
42.       Dia_Tm<-matrix(subset(base[,13],base[,3]==z & base[,16]==Mes & base[,14]!=4 & base[,17]<year), nrow=48)
43.       jm<-Dia_Tm[i,][Dia_Tm[i,]>0]
44.       bm<-1800/jm
45.       #Simulação chamadas abandonadas#
46.       Dia_ab1<-matrix(subset(base[,9],base[,3]==z & base[,16]==Mes & base[,14]!=4 & base[,17]<year), nrow=48)
47.       ab1<-Dia_ab1[i,][Dia_ab1[i,]>0]
48.       Dia_ab2<-matrix(subset(base[,10],base[,3]==z & base[,16]==Mes & base[,14]!=4 & base[,17]<year),
nrow=48)
49.       ab2<-Dia_ab2[i,][Dia_ab2[i,]>0]
50.       abandonadas<-ab1+ab2
51.       u<-which(abandonadas<=quantile(abandonadas,0.75)+1.5*(quantile(abandonadas,0.75)-
quantile(abandonadas,0.25)))
52.       ab<-abandonadas[u]
53.       s<-max(ab)
54.       if (s==0||s==1){a_b<-rltriangle(simu,1,1)*runif(simu,0,1)} else {a_b<-rltriangle(simu,1,s)}
55.       #####AJUSTE ENTRADA DE CHAMADAS#####
56.       #####Ajuste pela Poisson#####
57.       fit_p_m<-summary(fitdist(m,"pois", method="mme", gof="AD"))
58.       chi_p_m<-chisq.test(m,rpois(length(m), fit_p_m$estimate[1]))$p.value
59.       fitp_m<-matrix(c(1, fit_p_m$estimate[1], 0, fit_p_m$aic, fit_p_m$bic, chi_p_m))
60.       #####Ajuste pela Exponencial#####
61.       fit_e_m<-summary(fitdist(m,"exp", method="mme", gof="AD"))
62.       chi_e_m<-chisq.test(m,rexp(length(m), fit_e_m$estimate[1]))$p.value
63.       fite_m<-matrix(c(2, fit_e_m$estimate[1], 0, fit_e_m$aic, fit_e_m$bic, chi_e_m))
64.       #####Ajuste pela Normal#####
65.       fit_n_m<-summary(fitdist(m,"norm", method="mme", gof="AD"))
66.       chi_n_m<-chisq.test(m,rnorm(length(m), fit_n_m$estimate[1],fit_n_m$estimate[2]))$p.value
67.       fitn_m<-matrix(c(3, fit_n_m$estimate[1], fit_n_m$estimate[2], fit_n_m$aic, fit_n_m$bic, chi_n_m))
68.       #####Ajuste pela Gama#####
69.       fit_g_m<-summary(fitdist(m,"gamma", method="mme", gof="AD"))
70.       chi_g_m<-chisq.test(m,rgamma(length(m), fit_g_m$estimate[1],fit_g_m$estimate[2]))$p.value
71.       fitg_m<-matrix(c(4, fit_g_m$estimate[1], fit_g_m$estimate[2], fit_g_m$aic, fit_g_m$bic, chi_g_m))
72.       #####Ajuste pela Binomial Negativa#####
73.       fit_b_m<-summary(fitdist(m,"nbinom", method="mme", gof="AD"))
74.       chi_b_m<-chisq.test(m,rnbinom(length(m), mu=fit_b_m$estimate[2], size=fit_b_m$estimate[1]))$p.value
75.       fitb_m<-matrix(c(5, fit_b_m$estimate[1], fit_b_m$estimate[2], fit_b_m$aic, fit_b_m$bic, chi_b_m))
76.       Matriz_decisaom<-cbind(fitp_m,fite_m,fitn_m,fitg_m,fitb_m)
77.       Matriz_decisaom #### Matriz com as distribuições e seus parâmetros ####
78.       cm<-which(Matriz_decisaom[6,]>=0.05 & is.na(Matriz_decisaom[4,])==FALSE) #####Escolhe somente as
curvas que convergem####
79.       km<-which(Matriz_decisaom[4,]==min(Matriz_decisaom[4,cm])) #####Escolhe a melhor curva pelo teste
AIC####
80.       if (km==1){
81.         Taxac_m=fit_p_m$estimate[1]} else if (km==2)
82.         {Taxac_m=1/fit_e_m$estimate[1]} else if (km==3)
83.         {Taxac_m=fit_n_m$estimate[1]} else if (km==4)
84.         {Taxac_m=fit_g_m$estimate[1]/fit_g_m$estimate[2]} else
85.         {Taxac_m=fit_b_m$estimate[2]}
86.       Txc<-Taxac_m
87.       if (km==1){
88.         VarTn=fit_p_m$estimate[1]} else if (km==2)

```

```

89.   {VarTn=1/((fit_e_m$estimate[1])^2)} else if (km==3)
90.   {VarTn=(fit_n_m$estimate[2])^2} else if (km==4)
91.   {VarTn=fit_g_m$estimate[1]/((fit_g_m$estimate[2])^2)} else
92.   {VarTn=fit_b_m$estimate[2]+((fit_b_m$estimate[2])^2)/fit_b_m$estimate[1]}
93.   VarTn

94.   #####AJUSTE TEMPO MÉDIO DE ATENDIMENTO#####
95.   #####Ajuste pela Weibull#####
96.   fit_w_tm<-summary(fitdist(bm,"weibull", method="mle"))
97.   chi_w_tm<-chisq.test(bm,rweibull(length(bm), fit_w_tm$estimate[1], fit_w_tm$estimate[2]))$p.value
98.   fitwtm<-matrix(c(1, fit_w_tm$estimate[1], fit_w_tm$estimate[2], fit_w_tm$aic, fit_w_tm$bic, chi_w_tm))
99.   #####Ajuste pela Exponencial#####
100.  fit_e_tm<-summary(fitdist(bm,"exp", method="mme", gof="AD"))
101.  chi_e_tm<-chisq.test(bm,rexponential(length(bm), fit_e_tm$estimate[1]))$p.value
102.  fitetm<-matrix(c(2, fit_e_tm$estimate[1], 0, fit_e_tm$aic, fit_e_tm$bic, chi_e_tm))
103.  #####Ajuste pela Normal#####
104.  fit_n_tm<-summary(fitdist(bm,"norm", method="mme", gof="AD"))
105.  chi_n_tm<-chisq.test(bm,rnorm(length(bm), fit_n_tm$estimate[1], fit_n_tm$estimate[2]))$p.value
106.  fitntm<-matrix(c(3, fit_n_tm$estimate[1], fit_n_tm$estimate[2], fit_n_tm$aic, fit_n_tm$bic, chi_n_tm))
107.  #####Ajuste pela Gama#####
108.  fit_g_tm<-summary(fitdist(bm,"gamma", method="mme", gof="AD"))
109.  chi_g_tm<-chisq.test(bm,rgamma(length(bm), fit_g_tm$estimate[1], fit_g_tm$estimate[2]))$p.value
110.  fitgtm<-matrix(c(4, fit_g_tm$estimate[1], fit_g_tm$estimate[2], fit_g_tm$aic, fit_g_tm$bic, chi_g_tm))
111.  #####Ajuste pela Log Normal#####
112.  fit_lnm<-summary(fitdist(bm,"lnorm", method="mme", gof="AD"))
113.  chi_lnm<-chisq.test(bm,rlnorm(length(bm), fit_lnm$estimate[1], fit_lnm$estimate[2]))$p.value
114.  fitlnm<-matrix(c(5, fit_lnm$estimate[1], fit_lnm$estimate[2], fit_lnm$aic, fit_lnm$bic, chi_lnm))
115.  #####Matriz Estimativas#####
116.  Matriz_decisao_Tm<-cbind(fitwtm,fitetm,fitntm,fitgtm,fitlnm)
117.  Matriz_decisao_Tm #### Matriz com as distribuições e seus parâmetros ####
118.  ctm<-which(Matriz_decisao_Tm[,]>=0.05 & is.na(Matriz_decisao_Tm[,])==FALSE) ####Escolhe somente as
    curvas que convergem####
119.  ktm<-which(Matriz_decisao_Tm[,]==min(Matriz_decisao_Tm[,ctm])) ####Escolhe a melhor curva pelo teste
    AIC####
120.  if (ktm==1){
121.    TMAm=fit_w_tm$estimate[2]} else if (ktm==2)
122.    {TMAm=1/fit_e_tm$estimate[1]} else if (ktm==3)
123.    {TMAm=fit_n_tm$estimate[1]} else if (ktm==4)
124.    {TMAm=fit_g_tm$estimate[1]/fit_g_tm$estimate[2]} else
125.    {TMAm=exp(fit_lnm$estimate[1]+((fit_lnm$estimate[2])^2)/2)}
126.    TM<-TMAm
127.    if (ktm==1){
128.      VarSn=fit_w_tm$estimate[2]^2+(gamma(1+2/fit_w_tm$estimate[1])+(gamma(1+1/fit_w_tm$estimate[1]))^2)}
    else if (ktm==2)
129.      {VarSn=1/((fit_e_tm$estimate[1])^2)} else if (ktm==3)
130.      {VarSn=(fit_n_tm$estimate[2])^2} else if (ktm==4)
131.      {VarSn=fit_g_tm$estimate[1]/((fit_g_tm$estimate[2])^2)} else
132.      {VarSn=exp(2*fit_lnm$estimate[1]+(fit_lnm$estimate[2])^2)+(exp(fit_lnm$estimate[2]^2)-1)}
133.      VarSn

134.  #####Chamadas abandonadas#####
135.  fit_p_ab<-summary(fitdist(a_b,"pois", method="mme", gof="AD"))
136.  chi_p_ab<-chisq.test(a_b,rpois(length(a_b), fit_p_ab$estimate[1]))$p.value
137.  fitp_ab<-matrix(c(1, fit_p_ab$estimate[1], 0, fit_p_ab$aic, fit_p_ab$bic, chi_p_ab))
138.  #####Ajuste pela Exponencial#####
139.  fit_e_ab<-summary(fitdist(a_b,"exp", method="mme", gof="AD"))
140.  chi_e_ab<-chisq.test(a_b,rexponential(length(a_b), fit_e_ab$estimate[1]))$p.value

```

```

141. fite_ab<-matrix(c(2, fit_e_ab$estimate[1], 0, fit_e_ab$saic, fit_e_ab$bic, chi_e_ab))
142. #####Ajuste pela Normal#####
143. fit_n_ab<-summary(fitdist(a_b,"norm", method="mme", gof="AD"))
144. chi_n_ab<-chisq.test(a_b,rnorm(length(a_b), fit_n_ab$estimate[1],fit_n_ab$estimate[2]))$p.value
145. fitn_ab<-matrix(c(3, fit_n_ab$estimate[1], fit_n_ab$estimate[2], fit_n_ab$saic, fit_n_ab$bic, chi_n_ab))
146. #####Ajuste pela Gama#####
147. fit_g_ab<-summary(fitdist(a_b,"gamma", method="mme", gof="AD"))
148. chi_g_ab<-chisq.test(a_b,rgamma(length(a_b), fit_g_ab$estimate[1],fit_g_ab$estimate[2]))$p.value
149. fitg_ab<-matrix(c(4, fit_g_ab$estimate[1], fit_g_ab$estimate[2], fit_g_ab$saic, fit_g_ab$bic, chi_g_ab))
150. #####Ajuste pela Binomial Negativa#####
151. fit_b_ab<-summary(fitdist(a_b,"nbinom", method="mme", gof="AD"))
152. chi_b_ab<-chisq.test(a_b,rnbinom(length(a_b), mu=fit_b_ab$estimate[2],
size=fit_b_ab$estimate[1]))$p.value
153. fitb_ab<-matrix(c(5, fit_b_ab$estimate[1], fit_b_ab$estimate[2], fit_b_ab$saic, fit_b_ab$bic, chi_b_ab))
154. Matriz_decisao_ab<-cbind(fitp_ab,fite_ab,fitn_ab,fitg_ab,fitb_ab)
155. Matriz_decisao_ab #### Matriz com as distribuições e seus parâmetros ####
156. c_ab<-which(Matriz_decisao_ab[6,]>=0.05 & is.na(Matriz_decisao_ab[4,])==FALSE) ####Escolhe somente as
curvas que convergem####
157. k_ab<-which(Matriz_decisao_ab[4,]==min(Matriz_decisao_ab[4,c_ab])) ####Escolhe a melhor curva pelo
teste AIC###
158. if (k_ab==1){
159. Taxa_ab=fit_p_ab$estimate[1]} else if (k_ab==2)
160. {Taxa_ab=1/fit_e_ab$estimate[1]} else if (k_ab==3)
161. {Taxa_ab=fit_n_ab$estimate[1]} else if (k_ab==4)
162. {Taxa_ab=fit_g_ab$estimate[1]/fit_g_ab$estimate[2]} else
163. {Taxa_ab=fit_b_ab$estimate[2]}
164. Abandonada[z,i]<-ceiling(Taxa_ab)
165. ##### Dimensionamento Cemig-d #####
166. #Número inicial de atendentes por intervalo#
167. S<-round((Txc/TM)+0.5)
168. #Medidas de performance e variabilidade da fila G/G/s#
169. Cs<-VarSn/(TM^2)
170. Ca<-VarTn/(Txc^2)
171. Wq=(( (Txc/(TM*S))^2)*Cs+Ca)/(2*Txc*(1-Txc/(TM*S)))
172. Lq=Txc*Wq
173. while(Lq>(Txc*y)){
174. S=S+1
175. Wq=(( (Txc/(TM*S))^2)*Cs+Ca)/(2*Txc*(1-Txc/(TM*S)))
176. Lq=Txc*Wq
177. }
178. #Controle das chamadas abandonadas#
179. R=Txc/TM
180. Beta=(S-R)/sqrt(R)
181. T=30/1800
182. X=Beta*sqrt(TM/ceiling(Taxa_ab))
183. Y=sqrt(S*TM*ceiling(Taxa_ab))*T
184. X1=X+sqrt(ceiling(Taxa_ab)/(S*TM))
185. psi<-function(x,y){dnorm(x)/(1-pnorm(x+y))}
186. raz<-psi(X,Y)/psi(X1,Y)
187. prob<-1-(raz*exp(ceiling(Taxa_ab)*T))
188. teste<-round(prob,2)
189. Prob_aban[z,i]<-teste
190. if (teste <= IAb||is.na(teste)){S<-S
191. } else{repeat{
192. S=S+1
193. Beta=(S-R)/sqrt(R)
194. T=30/1800

```

```

195. X=Beta*sqrt(TM/ceiling(Taxa_ab))
196. Y=sqrt(S*TM*ceiling(Taxa_ab))*T
197. X1=X+sqrt(ceiling(Taxa_ab)/(S*TM))
198. psi<-function(x,y){dnorm(x)/(1-pnorm(x+y))}
199. raz<-psi(X,Y)/psi(X1,Y)
200. prob<-1-(raz*exp(ceiling(Taxa_ab)*T))
201. teste2<-round(prob,2)
202. if (teste2<=IAb||is.na(teste2)){
      a.      break
203. }
204. }}
205. Agentes[z,i]<-S
206. }
207. }
208. #Matriz de dimensionamento#
209. Agentes
210. }

```

A título de comparação com o modelo de dimensionamento proposto, implementou-se, no R-Studio, o programa *Turbo Table*, com predição do volume de chamada e estimativa do tempo médio de atendimento pela média dos intervalos de 30 minutos e dias de atividade. Nessa perspectiva, é coletado o banco de dados com as informações do volume e TMA estimados de ligações recebidas por dia, intervalo e mês de planejamento. Para executar o dimensionamento das equipes utilizando essa ferramenta, é necessário informar o mês desejado mês (*Mo*) e o Nível de Serviço (*SL*). A variável *SL* tem grandeza decimal e está compreendida entre 0 e 1.

```

1. dimensionamento_TT=function(Mo,SL){
2.   require(queueing)
3.   bd<-read.csv("C:\\Back up_Alexandre\\Mestrado\\Produção\\Dissertação\\Bancos de
   Dados\\chamadas_jan.csv", header = TRUE, sep = ";")
4.   Mes<-Mo
5.   INS<-SL
6.   #bd#
7.   #transformando o bd em matriz#
8.   base<-array(bd)
9.   require(fitdistrplus)
10.  require(gmp)
11.  require(triangle)
12.  AgentesM<-matrix(nrow=7,ncol=48)
13.  for (k in 1:7){
14.    for (i in 1:48){
15.      Dia_c<-subset(base[,4],base[,2]==k & base[,1]==Mes & base[,3]==i)
16.      Dia_T<-subset(base[,5],base[,2]==k & base[,1]==Mes & base[,3]==i)
17.      Taxac<-Dia_c
18.      TMA<-Dia_T
19.      TaxaT<-1800/TMA
20.      Erlangs<-round((Taxac*TMA/1800)+0.5)
21.      r=Taxac/TaxaT #TrafficRate
22.      NAgents=0
23.      if(Erlangs < 1) {NAgents=1} else {NAgents=round(Erlangs)}

```

```

24.      Utilisation=r/NAgents
25.      while(Utilisation >= 1){
26.          NAgents=NAgents+1
27.          Utilisation=r/NAgents
28.      }
29.      SLA=INS
30.      MaxIterate=NAgents*10
31.      MaxAccuracy=0.00001
32.      Server=0
33.      c=0
34.      st=30/1800
35.      for (count in 1:MaxIterate){
36.          Utilisation=r/NAgents
37.          if(Utilisation < 1){
38.              Server=NAgents
39.              c<-C_erlang(Server,r)
40.              SLQueued=1-c*exp((r-Server)*st/TMA)
41.              if(SLQueued<0){SLQueued=0}
42.              if(SLQueued>=SLA){count=MaxIterate}
43.              if(SLQueued>(1-MaxAccuracy)){count=MaxIterate}
44.          }
45.          if(count!=MaxIterate){NAgents=NAgents+1}
46.      }
47.      AgentesM[k,i]<-NAgents
48.  }
49.  }
50.  AgentesM
51.  }

```