UNIVERSIDADE DE BRASÍLIA

FACULDADE DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA ELÉTRICA

ANÁLISE E PROJEÇÃO DE TRÁFEGO TELEFÔNICO

TARCISIO DE NEGREIROS BOMFIM

ORIENTADOR: JOÃO MELLO DA SILVA

DISSERTAÇÃO DE MESTRADO EM ENGENHARIA ELÉTRICA

PUBLICAÇÃO: PPGENE.DM – 083/2011

BRASÍLIA / DF: 12/2011

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

ANÁLISE E PROJEÇÃO DE TRÁFEGO TELEFONICO

TARCISIO DE NEGREIROS BOMFIM

DISSERTAÇÃO DE MESTRADO SUBMETIDA AO DEPARTAMENTO DE ENGENHARIA ELÉTRICA DA FACULDADE DE TECNOLOGIA DA UNIVERSIDADE DE BRASÍLIA, COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE PROFISSIONAL EM INFORMÁTICA FORENSE E SEGURANÇA DA INFORMAÇÃO.

APROVADA POR:

JOÃO MELLO DA SILVA, UnB
(ORIENTADOR)

LUIS FERNANDO RAMOS MOLINARO, Doutor, UnB
(EXAMINADOR INTERNO)

JOÃO CARLOS FELIX DE SOUZA, Doutor,Unb
(EXAMINADOR EXTERNO)

DATA: BRASÍLIA/DF, 30 DE DEZEMBRO DE 2011.

FICHA CATALOGRÁFICA

BOMFIM, TARCISIO DE NEGREIROS
Análise e projeção mensal de tráfego telefônico. [Distrito Federal] 2011.
xiv, 97p., 210 x 297 mm (ENE/FT/UnB, Mestre, Dissertação de Mestrado - Engenharia -
Universidade de Brasília. Faculdade de Tecnologia.
Departamento de Engenharia Elétrica.
1. Tráfego Telefônico 2. Análise de Tráfego telefônico
3. Projeção de tráfego
I. ENE/FT/UnB. II. Título (Série)

REFERÊNCIA BIBLIOGRÁFICA

BOMFIM, T. N. (2011). Análise e projeção de tráfego telefônico. Dissertação de Mestrado, Publicação PPGENE.DM - 083/11, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 97p.

CESSÃO DE DIREITOS

NOME DO AUTOR: Tarcisio de Negreiros Bomfim

TÍTULO DA DISSERTAÇÃO: Análise e projeção de tráfego telefônico.

GRAU/ANO: Mestre/2011.

Tarcisio de Negreiros Bomfim
CLSW 102 Bloco C Ed. Atlantis
CEP 33423-179 Brasília DF - Brasil

Dedico à Silvia Helena, minha mulher, e aos meus filhos Cristina, Carolina e Victor.

AGRADECIMENTOS

Ao meu orientador Prof. João Mello, por acreditar em mim e pelo constante apoio, incentivo, dedicação e amizade essenciais para o desenvolvimento deste trabalho.

Ao Prof. Flávio Elias, do Curso de Engenharia de Redes de Comunicação - Departamento de Engenharia Elétrica, pelo apoio e orientação na elaboração deste trabalho.

A todos, que colaboraram e me apoiaram nesta caminhada, os meus sinceros agradecimentos.

RESUMO
ANÁLISE E PROJEÇÃO DE TRÁFEGO TELEFÔNICO

Autor: Tarcisio de Negreiros Bomfim
Orientador: João Mello da Silva
Programa de Pós-graduação em Engenharia Elétrica
Brasília, Dezembro de 2011

Apresentar uma solução para analisar o tráfego telefônico mensal com o objetivo de descobrir anomalias no seu comportamento e para projetar o tráfego do mês corrente com base no tráfego dos sete primeiros dias de cada mês, estimou o tráfego para todo o mês e a receita associada a este tráfego. Através da análise e projeção mensal é possível detectar desvios e antecipar ações de engenharia ou de marketing para minimizar os impactos na receita operacional de voz. Este processo de análise diária do tráfego é um diferencial de mercado para Empresas de Telecom porque aumenta e torna mais rápida a sua capacidade de reação à concorrência e capacidade de recuperação junto aos órgãos reguladores**.**

ABSTRACT

ANALYSIS AND FORECAST OF TELEPHONE TRAFFIC

Author: Tarcisio de Negreiros Bomfim

Supervisor: João Mello da Silva

Programa de Pós-graduação em Engenharia Elétrica

Brasília, December of 2011

Build a process to analyse telephone traffic each month with the aim of discovering anomalies in their behavior and to forecast the current month traffic based on the first seven days . Estimate the traffic for the entire month together with associated revenue . Through analysis and monthly forecast is possible to detect deviations and anticipate actions in engineering or marketing to minimize impacts on voice revenue. This process of daily traffic analysis is a market differentiator for Telecom companies as it increases speed and their ability to fight back against the competition and resilience with regulatory agencies

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE SÍMBOLOS, NOMENCLATURA E ABREVIAÇÕES

BL	Banda Larga
BPM	Business Process Management
BI	Business Inteligence
CDR	CALL DATA RECORD ou CALL DETAIL RECORD
DIA ÚTIL	Dia de uma semana de trabalho de segunda à sexta-feira
Gb	Gigabits
H.M.M.	Hora de maior movimento, é o intervalo de hora de maior tráfego
Mb	Megabits
PME	Pequena e Média Empresa
SAS	Statiscal Analysis System é um sistema com pacote de análise estatística
SCM	Serviço de Comunicação Multimidia
SMS	Short Menssages Service
UIT	União Internacional de Telecomunicações

1. INTRODUÇÃO

Um dos grandes problemas das operadoras de Telecom no Brasil e no Mundo é o gerenciamento da receita voz. O serviço de voz não demanda grandes investimentos porque os processos relacionados a operação do serviço estão totalmente implantados e a tecnologia já está completamente dominada. Entretanto como a receita produzida pelo serviço de voz ainda tem bastante peso no resultado econômico-financeiro das empresas existe a necessidade de um acompanhamento constante do tráfego de voz. Qualquer variação no tráfego causa impacto na receita de voz.

A seguir, dois casos concretos são apresentados para ilustrar o impacto de desvios do tráfego de voz no resultado mensal de uma Operadora de Telecom. A necessidade de a empresa gerenciar o impacto desse desvio foi a inspiração para esta dissertação. É importante comentar que a receita de voz das grandes empresas de Telecom, ainda é um item de grande relevância no resultado financeiro destas empresas requerendo uma gestão continua, pois qualquer desvio representa um impacto de bilhões de reais. Na tabela 1.1 observa-se a receita de voz das empresas Telefonica e Oi no 1º. Trimestre de 2011.

Tabela 1.1 Receita Operacional da Telefonica[1] e da Oi[2] no 1º. Trimestre 2011

Item	Receita (R$ milhões)	
	Telefonica	Oi
Operacional Bruta	**R$ 5.388,90**	**R$ 8.221,30**
Telefonia Fixa	**R$ 3.766,20**	**R$ 5.158,40**
Serviço Local	R$ 2.385,80	R$ 3.878,70
Serviço Longa Distância	R$ 1.380,40	R$ 1.279,70
Dados	**R$ 1.122,30**	**R$ 2.277,30**
Outros	**R$ 500,40**	**R$ 785,60**
Receita de Voz[3]	**R$ 3.761,52**	**R$ 5.150,95**
Linhas em Serviços(*)	**15.602**	**24.603**

(*) inclui BL e TV

[1] Relatório: Resultados Janeiro Março / 2011 da Telefonica de 12 de Maio de 2011
[2] Relatório Trimestral: Relação com Investidores 1T2011 da Oi de 28 de Abril de 2011
[3] A Receita de voz foi estimada com base nas informações da receita de assinatura informadas nos relatórios

1.1. MOTIVAÇÃO E JUSTIFICATIVA

Em meados de Junho de 2007, durante a análise estimada do resultado mensal de uma Operadora de Telecom brasileira foi constatado um desvio da ordem de 10% no valor mensal da receita de voz inter-redes[4]. A análise desta receita é realizada com base no tráfego apurado na 1ª. semana do mês corrente e em comparação com o orçamento mensal elaborado no inicio do ano. Neste caso não foi possível identificar, ainda em Junho, os fatos relevantes responsáveis pelo desvio da receita e decidiu-se aguardar o fechamento mensal da receita do mês para dar prosseguimento ao processo com uma análise mais detalhada do desvio do tráfego. A análise do resultado de Junho, realizada durante o mês de Julho revelou uma falha no processo de faturamento dos bilhetes de chamada em várias filiais. O diagnóstico completo da falha foi confirmado no final de Julho e somente em Agosto foram disparadas ações para corrigir a falha ocorrida em Junho. A receita de inter-redes ficou reduzida na ordem de 10% durante o período de três meses. Este prejuízo poderia ter sido reduzido caso a Empresa dispusesse de um processo estruturado e ágil para antecipar a análise do desvio e da implantação das ações corretivas.

Em outra Operadora em Agosto de 2007 foi constatado um desvio da ordem de 5% a maior na despesa de interconexão local[5] de uma grande filial confirmava uma tendência crescente, iniciada dois meses antes. A análise do desvio revelou que uma empresa concorrente fora bem sucedida em uma ação agressiva de conquistar clientes de alto tráfego. A demora na obtenção de um diagnóstico confiável atrasou em três meses a ação de retenção e causou perdas financeiras relevantes nessa filial. Neste caso, para evitar o prejuízo a Empresa deveria dispor de um processo ágil e confiável para predizer desvios na despesa. Uma projeção do tráfego mensal poderia ter sido elaborado com base no tráfego dos primeiro dias, permitindo a antecipação do diagnostico que revelaria a ação da concorrência. A ação de retenção dos clientes poderia ser efetivada ainda em tempo de reduzir o prejuízo da filial

O produto Voz na Telefonia fixa já alcançou a maturidade no Mercado brasileiro e não demanda grandes investimentos, desta forma com pequenos e localizados

[4] Receita Inter-rede é a receita de Voz gerada por uma ligação com destino à rede de outro serviço da mesma operadora ou com destino à rede de outra operadora

[5] Despesa de Interconexão Local é o valor pago por uma operadora por ligações geradas por seus usuários e destinadas à rede de outra operadora

investimentos podem- om no mercado brasileiro em Novembro de 2006. A Operadora de Telecom fez um contrato com uma empresa de TV aberta e com lojas de varejo. Com base nesse contrato a Operadora de

chamadas. As lojas disponibilizavam prêmios (de celulares, computadores, televisores e games) através de um leilão onde os lances eram direcionados para o número disponibilizado. Os telespectadores davam lances com detalhes de centavos. O lance único era contemplado independente do valor. O leilão era realizado no horário após 12 horas da noite, no período até as 5 horas da manhã. Eram feitos mais de leilão por noite com duração de uma hora em média. A quantidade de ligações e o tráfego gerado por elas causaram um impacto relevante na receita mensal de voz da Operadora. Este projeto foi um sucesso pela relação beneficio/custo, sendo logo implantado pelas outras Operadoras.

Outro exemplo é programa Big Brother Brasil 11 que a cada paredão[6] gera um número da ordem de R$ 50 milhões de votos, dos quais, 10% são registrados por telefone (Globo, 2011). Considerando que a maioria dos que votam por telefone, fazem-no através de celulares e utilizando a tarifa de uma ligação VC1 de R$ 0,78950 de acordo com Telefonica (2011), o valor faturado por paredão é de R$ 3,95 milhões[7]. Durante as 11 semanas de duração do programa, aconteceram 10 votações, conclui-se que o valor faturado por programa somente com ligações telefônicas é de R$ 39,5 milhões

Devido sua alta relevância, a receita de voz necessita de um gerenciamento constante para evitar ou reagir aos ataques da concorrência cada vez mais freqüentes e diversificados.

[6] Paredão é nome do processo de votação, onde os telespectadores indicam quem deve ser eliminado através de votação pela Internet, SMS ou ligações telefônicas
[7] Valor faturado = 10% x 50 milhões x R$ 0,78950

1.2. OBJETIVOS

Os exemplos apresentados no item 1.1 mostram a necessidade de implantação de processo prático, ágil e confiável de análise e predição com o objetivo de apontar tendências no tráfego cursado[8].

O contexto de como o trabalho de análise e estimativa de tráfego é realizado nas operadoras de Telecom aponta para necessidade de criar um ambiente de Business Inteligence BI. Este ambiente dá condições às áreas de Análise e Planejamento da receita de voz para utilizar tecnologias e processos que tratam os dados com o objetivo de obter informações necessárias para entender e analisar o desempenho do negócio , a receita de voz (Davenport, 2007, p.7). As opções de ferramentas para suportar o processo de gerenciamento analítico segundo Davenport (2007, p. 8) são muitas, desde simples as mais simples como o Excel, passando por soluções de pacotes de software estatísticos como o Minitab e atingido o máximo com conjuntos de soluções de business intelligence como o SAS, Cognos, BussinessObjects, nas aplicações industriais de predição como o Fair Issac e em grandes sistemas empresariais como SAP e Oracle .

Dentre as ferramentas disponíveis no mercado, a escolhida foi o Statiscal Analysis System-SAS. A utilização do SAS tornou o processo mais ágil, aumentou sua confiabilidade, além de possibilitar a sua utilização por profissionais de diversas áreas como financeira, planejamento e controle sem necessidade de saber programar o código da SAS (Schlotzhauer, 2009).

1.3. METODOLOGIA E ORGANIZAÇÃO DA DISSERTAÇÃO

O processo de análise e projeção desenvolvido começa pela preparação de uma base de dados com informações históricas do tráfego diário distribuídos por filial que são formadas por cidades e estados. Em cada cidade o tráfego é medido nas centrais telefônicas que são instaladas nos bairros. Cada chamada originada em uma central[9] é registrada através de um CDR que de acordo com ITU-T (1998) e ITU-T (2001) contém todas as informações necessárias ao o gerenciamento e tarifação de cada chamada. As principais informações são o número de origem da chamada, hora, minuto e segundo de cada

[8] Tráfego cursado é tráfego efetivo, resultante de ligações telefônicas completadas
[9] Chamada originada por um cliente que é atendido pela central do bairro

chamada, duração, degrau, e tipo de ligação. A preparação da base histórica é atividade simples na forma, mas complexa na sua estruturação devido ao alto volume de informações coletadas. Na dissertação foi utilizada uma amostra com o tráfego de uma cidade com mais de 1 milhão de habitantes onde cada arquivo de tráfego diário tem em média 6 milhões de CDR e ocupa 350Mb de memória, ou seja para cada mês de tráfego é necessário processar uma base com mais de 10 Gb de memória.

A partir da base estruturada, foi analisado o comportamento diário do tráfego para identificar tendências recorrentes como aquela em que o tráfego no final do mês é maior que no início. Outra tendência observável é o tráfego nos fins de semana, bem menor do que o de um dia útil na maior parte das centrais telefônicas. Entretanto, em centrais localizadas em área de veraneio o tráfego no fim de semana é maior que o observável durante a semana. Existe também a tendência das horas de maior movimento, onde o tráfego cursado nos períodos de 9 às 11horas, 14 às 16 horas e 20 às 22horas é bem maior que nos outros horários (Tude, 2003). As tendências escolhidas serão uma referência para análise das informações recebidas a cada dia e um vetor para as projeções serão elaboradas.

Considerando a agilidade como uma prioridade para o processo de gerenciamento do tráfego relacionado à receita de voz, fica claro que o mais importante é começar a análise do tráfego de uma forma mais geral, ou seja, com a visão de filial e depois dentro de cada filial, onde for necessário, abrir a análise por cidade, por central, por hora podendo até chegar ao nível de chamada. Inicialmente, com base histórica estruturada, são definidos os valores representativos de tráfego diários para cada filial. Diferentes filiais podem ter diferentes valores representativos. Em uma determinada filial, é possível definir um valor que represente o tráfego de um dia útil e outro que represente os demais dias da semana. O tráfego mensal estimado para esta filial é obtido a partir da quantidade de dias úteis e demais dias multiplicada pelo tráfego representativo de cada um dos tipos de dias respectivamente. Entretanto para outras filiais, a alternativa é definir um valor para cada dia semana, considerando que o tráfego de uma segunda-feira pode ser diferente do tráfego da terça e dos outros dias da semana, assim como, no fim de semana o tráfego dos domingos podem ter características e volume bem diferentes dos sábados.

Na segunda etapa da análise, utiliza-se o tráfego dos primeiros dias do mês de cada filial para comparar com os valores representativos de cada dia de cada filial. Quando os novos valores forem comparáveis aos valores representativos dentro de um determinado

intervalo, a base histórica é atualizada com estes novos valores de tráfego e os valores representativos são recalculados. Estes são utilizados para estimar o tráfego mensal. No capítulo 6 este processo será melhor explicado e detalhado. Quando algum novo valor não for comparável ao valor representativo dentro de um determinado intervalo, a base histórica não será atualizada. O tráfego mensal será estimado com nos novos valores que forem coerentes. Além disso, o processo prevê a coleta de mais uma amostra com sete dias da filial em questão e nova comparação com a base histórica. Caso esta nova análise confirme a nova tendência do tráfego, esta será considerada na estimativa do tráfego mês no lugar da base histórica.

Na etapa seguinte é feita uma projeção do tráfego mensal a partir da base histórica atualizada e com base na quantidade de dias úteis, sábado, domingo e feriados no mês. Uma vez estimado o tráfego cursado da empresa, resultado do somatório de todas filiais, converte-se o tráfego em receita utilizando uma relação histórica de tráfego/receita. A idéia de transformar o tráfego em receita é utilizar uma linguagem mais conhecida nas empresas. Todos conhecem bem a receita mensal da empresa, mas poucos conhecem o tráfego cursado.

Na dissertação é apresentada uma demonstração da utilização do SAS em toda a cadeia do processo de estimativa do tráfego, desde a estruturação de uma base histórica de tráfego, passando pelas etapas de seleção dos campos de interesse, pela atualização do histórico, de alerta para possíveis problemas na coleta dos dados e estimativa do tráfego mensal. Resumindo o roteiro da dissertação é o seguinte: uma introdução; preparação da base histórica; análise do tráfego; atualização da base; estimativa mensal e conclusões.

É importante ressaltar que o objetivo principal desta dissertação é apresentar uma proposta de solução para implantação de um processo que atenda a demanda de das operadoras de telecomunicação de ter um processo confiável e ágil para estimar a receita operacional mensal de voz, conciliando a experiência e cultura de cada empresa com as melhores práticas de análise e projeção de tráfego, com a utilização de ferramentas atualizadas e poderosas como o SAS

2. ESTADO DA ARTE

Neste capítulo mostra-se o cenário de como as operadoras de Telecom lidam com a necessidade de analisar o histórico do tráfego e projetar o tráfego para o mês corrente e quais são as regulamentações nacionais e internacionais seguidas. Os parâmetros de estatística e conceitos utilizados pelas equipes responsáveis pelo trabalho de análise e estimativa de tráfego são apresentados, bem como o ambiente gerencial e processo utilizado em cada área. Mostra-se também as ferramentas de TI que são empregadas.

2.1. CENÁRIO ATUAL

Atualmente as operadoras de Telecom lidam com os processos de analise e estimativa de tráfego telefônico de uma forma diferenciada e individualizada. As empresas de menor porte concentram todo o processo numa só área, na maioria das vezes na área de TI, que recebe os arquivos com os CDRs e transforma-os em tráfego a ser tarifado, além de fazerem o gerenciamento e administração do mesmo. Quando a área financeira necessita de alguma análise ou estimativa, gera uma demanda para área de TI.

As operadoras de maior porte têm processos de gerenciamento de tráfego executados em diferentes áreas e algumas delas com superposição de atividades. Os dados de CDRs são gerados nas centrais telefônicas, e enviados para os mediadores que são equipamentos distribuídos em pontos estratégicos da rede para coletar os dados de todas as centrais em operação na planta da operadora seguindo as recomendações do ITU-T (1998). Os dados são normalmente distribuídos para área de Gerência da Rede, que é responsável por garantir a operação de todos os elementos de rede: central, rádio, roteadores, concentradores ou mediadores conforme as recomendações do ITU-T (2002). A área de TI também recebe os dados de CDRs, que são processados e geram arquivos com tráfego em minutos que são enviados à área financeira para acompanhamento do processo de tarifação que é realizado pela TI que se utiliza dos mesmos arquivos para executar o processo de tarifação e emissão de faturas. Existem ainda as áreas de análise e planejamento que recebem e trabalham o tráfego de forma setorizada: tráfego de origem fixa; tráfego de clientes corporativos; tráfego de clientes governo; tráfego de clientes pequenas e médias empresas - PME; tráfego de origem móvel e tráfego de dados. Estas áreas são as que têm o maior interesse no tráfego tarifado, que é a fonte geradora de uma das maiores receitas da

empresa, a receita operacional de voz. Elas utilizam o tráfego na elaboração do orçamento anual e no gerenciamento mensal do resultado financeiro, conforme modelo das tabelas 2.1 e 2.2.

Tabela 2.1 Análise da receita de voz de uma operadora de Telecom

Receita (R$ milhões)	Mês Atual			Desvio			
	Anterior	Atual	Orç mês	Anterior		Orçado	
	R$	R$	R$	R$	%	R$	%
Telefonia Fixa	1.277,6	1.255,4	1.260,6	-22	-2%	-5	0%
Serviço Local	802,6	795,3	801,6	-7	-1%	-6	-1%
Mensalidade	1,58	1,56	1,60	-0	-1%	-0	-3%
Voz	801,0	793,7	800,0	-7	-1%	-6	-1%
Serviço Longa distancia	475,0	460,1	459,0	-15	-3%	1	0%
Voz	475,0	460,1	458,0	-15	-3%	2	0%

Tabela 2.2 Análise do tráfego de voz de uma operadora de Telecom

Tráfego (milhões minutos)	Mês Atual			Desvio			
	Anterior	Atual	Orç mês	Anterior		Orçado	
					%		%
Telefonia Fixa	8.436,2	8.346,0	8.404,8	-90	-1%	-59	-1%
Serviço Local	7.834,5	7.763,2	7.824,7	-71	-1%	-62	-1%
Mensalidade	0,00	0,00	0,00	0		0	
Voz	7.834,5	7.763,2	7.824,7	-71	-1%	-62	-1%
Serviço Longa distancia	601,6	582,8	580,1	-19	-3%	3	0%
Voz	601,6	582,8	580,1	-19	-3%	3	0%

Nas tabelas 2.1 e 2.2, está representado um modelo de análise da receita e do tráfego com base nas melhores práticas de gestão utilizadas pelas operadoras de Telecom. A análise é feita em duas etapas, sendo a primeira uma comparação com o mês anterior, onde a receita de Telefonia Fixa de R$ 1.255,4 milhões no mês atual está 22 milhões menor que a receita do mês anterior que foi de R$ 1.277,6 milhões com um desvio[10] de -2%. Na segunda etapa, é feita uma comparação com o valor orçado para o mês, onde a receita de R$ 1.255,4 milhões está R$ 5 milhões menor que o orçamento que é de R$ 1.260,6 milhões. A mesma análise pode ser feita com relação ao tráfego. O tráfego do mês atual de 8.346,0 milhões de minutos está 90 milhões de minutos a menor que o tráfego do mês anterior que foi de 8.436,2 milhões de minutos com um desvio[11] de -1%. Da mesma

[10] O desvio é calculado pela divisão da diferença entre a receita do mês atual e a do mês anterior dividida pela receita do mês atual

[11] O desvio é calculado pela divisão da diferença entre o tráfego do mês atual e o do mês anterior dividida pelo tráfego do mês atual

forma, o tráfego do mês atual está 59 milhões de minutos menor que o valor orçado que é 8.404,8 milhões minutos.

Dentro do orçamento, as receitas de tráfego de voz são ainda as de maior relevância. As equipes responsáveis pela elaboração do orçamento são cada vez mais cobradas e trabalham sob grande pressão. Primeiro, elas são pressionadas a partir de Julho para elaboração do orçamento do ano seguinte, um trabalho que deve ser concluído em Novembro para ser ainda revisado e aprovado antes do final do ano. Segundo, elas são responsáveis durante todo o ano, várias vezes em cada mês para detectar e explicar desvios ocorridos no resultado em relação ao orçamento e aos anos anteriores. É possível imaginar que em um orçamento anual de dezenas de bilhões de reais, qualquer pequeno desvio percentual produz um impacto de centenas de milhões de reais, que afeta o resultado e

is ou três meses.

No cenário atual estas áreas têm uma carência muito grande de informações e meios para realizar seus trabalhos. A maioria delas recebe informações da área de TI em arquivos Excel, em planilhas já consolidadas e em formato previamente acordado. Fazer uma análise da receita de voz a partir do tráfego quando o comportamento segue o padrão não tem grandes dificuldades. Entretanto quando ocorrem desvios relevantes ou distorções no tráfego, existe a necessidade de uma análise mais especifica e localizada, ou seja uma análise do tipo top down[12](Celebroni, 2009). Nestes casos, os arquivos consolidados e as planilhas padronizadas não permitem que a equipe responsável pesquise a causa do desvio ou da distorção porque sua flexibilidade é limitada. Além disso a agilidade da análise fica comprometida porque qualquer alteração nos arquivos padronizados deve ser feita por TI.

Na tabela 2.3 é apresentado um exemplo de uma análise top down do tráfego de uma operadora aberto por quatro filiais. Na tabela, o desvio, aproximado, de 0% do tráfego de 8.36 bilhões de minutos da operadora no mês atual em comparação ao tráfego orçado de 8,39 bilhões de minutos mostra um cenário de aparente de normalidade. Entretanto, quando o tráfego é analisando por filial, percebem-se alguns pequenos desvios nas filiais A e C e uma grande distorção na filial com desvio de 10% com relação ao orçado.

[12] Análise top down é a análise de cima para baixo, termo utilizado em BPM para caracterizar um processo de análise onde o processo se inicia no nível mais geral e gradativamente desce para os níveis seguintes até a identificação do problema ou até mais baixo da análise

Tabela 2.3 Modelo de análise top down no tráfego de uma operadora de Telecom

Tráfego (milhões minutos)	Mês Atual			Desvio (%)			
	Anterior	Atual	Orç mês	Anterior		Orçado	
Operadora	8.436,2	8.356,5	8.394,8	-80	-1%	-38	0%
Filial A	3.350,0	3.300,0	3.260,0	-50	-1%	40	1%
Filial B	2.070,0	2.086,5	2.090,0	16	1%	-4	0%
Filial C	2.075,0	2.070,0	2.044,8	-5	0%	25	1%
Filial D	941,2	900,0	1.000,0	-41	-4%	-100	-10%

Este modelo explica porque é importante abrir a análise em filial para localizar onde ocorreu o grande desvio de 100 milhões de minutos. Identificando-se uma filial apenas, a análise prossegue para descobrir se o impacto aconteceu em uma ou mais cidades. A análise pode chegar ao nível de central ou no caso de desvio temporário, ao nível do intervalo de tempo quando ocorreu o impacto. Nestes casos as informações de tráfego previamente consolidadas não atendem à necessidade da equipe responsável. A equipe necessita de uma abertura que não está disponível e a área de TI que atende toda a empresa, não tem um processo estabelecido que entregue em tempo um relatório com as informações necessárias. A solução muitas vezes buscada e algumas atendida é dotar área de planejamento com uma estrutura própria de TI para garantir informações no tempo certo com a abertura desejada. Entretanto, esta estrutura não sobrevive à primeira reestruturação empresarial porque a tendência das empresas é de terceirização e a atividade de TI é uma das primeiras, transformadas em fábricas de TI terceirizadas.

Em resumo, no cenário atual as grandes operadoras como Telefonica/Vivo, Embratel/Claro/Net, Oi e TIM/Intelig, como as de médio porte GVT/Vivendi, Nextel e CTBC/Algar e como também as pequenas empresas prestadoras de Serviço de Comunicação Multimídia-SCM não têm uma estratégia bem definida do processo de análise e projeção do tráfego. Dessa forma este trabalho tem sido executado sem muita eficiência e com baixa eficácia deixando as operadoras sem um processo definido para o gerenciamento da receita de voz.

3. PREPARAÇÃO DA BASE DE DADOS

Neste capítulo apresenta-se a proposta de um processo para preparação uma base de dados gerencial para ser utilizada no processo de análise do tráfego telefônico. A idéia é estruturar uma base de dados acessível para a equipe de analistas de tráfego de uma forma ágil e abrangente

3.1. PROCESSO DE COLETA

A base necessária para o gerenciamento do tráfego associado à receita de voz é constituída por dados contidos nos CDRs que são gerados nas centrais telefônicas, coletados pro mediadores, transportados pela rede dados das operadoras de Telecom e armazenados em banco de dados.

Na figura 3.1 apresenta-se um resumo do processo coleta de CDR onde, para toda chamada ou tentativa de chamada a central de comutação gera um arquivo contendo todas as informações sobre o evento. Este arquivo chamado de CDR ou bilhete é armazenado em um banco de dados. Os CDR são analisados e todos

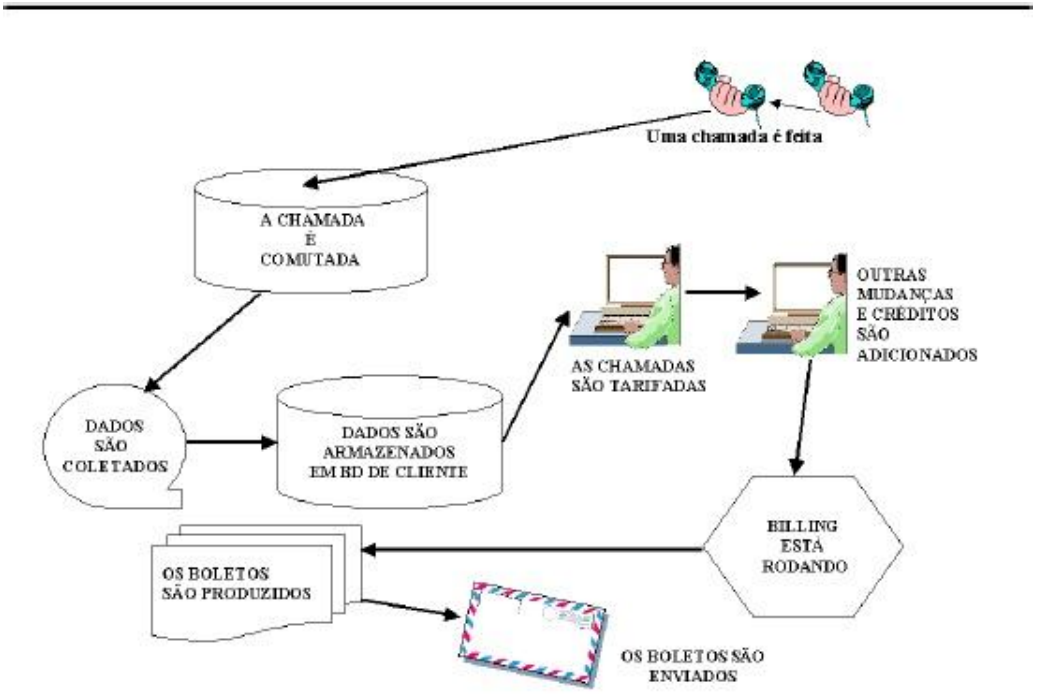


Figura 3.1 Processo de coleta de CDR.

aqueles completos, isto é, com todos os dados da chamada são devidamente contabilizados e encaminhados para Tarifação onde são transformados em faturas a serem enviadas para cobrança.

O CDR gerado por cada chamada telefônica é padronizado pela União Internacional de Telecomunicações (ITU-T, 1998) e aqui no Brasil utiliza 69 campos de dados como mostrado na tabela 3.1. Estes dados são uma amostra da base utilizada na dissertação e obtida de Costa (2010) por exportação através do SAS. Considerando as centenas de milhões de usuários de telefonia no país gerando milhões de chamadas por dia podemos imaginar a quantidade de informações a serem transmitidas, armazenadas e processadas.

Tabela 3.1 Relação de campos de um CDR

Campo 1	CENTRAL	AB1	AB1	CD2	CD3	EF2	EF3
Campo 2	RTRA_NUMERO_REMESSA	4869450	4869450	4869450	4869450	4869450	4869450
Campo 3	SEQUENCIAL	571	575	576	577	1536	1537
Campo 4	TELEFONE_ORIGEM	69199899994	6930283029	6930811014	6930925662	6999759064	6921013624
Campo 5	PARTE_TARIFARIA	1	0	0	0	0	0
Campo 6	CATEGORIA	1	1	1	1	1	1
Campo 7	TELEFONE_DESTINO	8946199737095	896499576124	896499525331	896484241735	8936134219	8936214323
Campo 8	FIM_SELECAO	1	1	1	1	1	1
Campo 9	HORA_INICIO	83714	91429	82958	80116	91705	91544
Campo 10	DURACAO	0,55	0,07	1,32	1,05	0,30	1,59
Campo 11	DIA_MES	01:11:00	01:11:00	01:11:00	01:11:00	01:11:00	01:11:00
Campo 12	CAUSA_SAIDA	0	0	0	0	0	0
Campo 13	CONTADOR_SAIDA	1	1	1	1	1	1
Campo 14	NUMERO_ROTA_ENTRADA	9700	IGW2	IGW1	IGW1	ICG1	IEBC
Campo 15	NUMERO_JUNTOR_ENTRADA	190	3	15	597	336	8
Campo 16	NUMERO_ROTA_SAIDA	IGW1	9700	9700	IMG1		
Campo 17	NUMERO_JUNTOR_SAIDA	849	11	20	45		
Campo 18	CLASSE_CHAMADA	10	10	10	10	10	10
Campo 19	CODIGO_EVENTO						
Campo 20	CONFIABILIDADE	0	0	0	0	0	0
Campo 21	CLASSE_TARIFA						
Campo 22	LOCA_ID_ORIGEM	611	4933	5001	1595	3045	3535
Campo 23	LOCA_ID_DESTINO	611	3011	1645	3535	3535	3535
Campo 24	PAIS_CODIGO_ORIGEM						
Campo 25	PAIS_CODIGO_DESTINO						
Campo 26	AECP_ID						
Campo 27	RDTR_MES_ANO_REF	11/2009	11/2009	11/2009	11/2009	11/2009	11/2009
Campo 28	IND_CHAMADA_NACIONAL	S	S	S	S	S	S
Campo 29	DEGRAU_TARIFARIO	2	2	2	2	1	
Campo 30	TIPO_TARIFA_VALORACAO	M	M	M	M	M	F
Campo 31	TIPO_TRAFEGO	5	5	5	5	4	0
Campo 32	TIPO_TERMINAL_ORIGEM	M	F	F	F	M	F
Campo 33	TIPO_TERMINAL_DESTINO	M	M	M	M	F	F
Campo 34	TELEFONE_DISCADO						
Campo 35	VALOR_CHAMADA						
Campo 36	TCTE_CODIGO	50	10	10	10	50	10
Campo 37	TSTR_CODIGO	33	33	33	33	27	25
Campo 38	DDDP_DDD_NUMERO_ORIGEM	61	62	62	62	63	64
Campo 39	DDDP_PREF_NUMERO_ORIGEM	9989	3028	3081	3092	9975	2101
Campo 40	CODIGO_SELECAO_NACIONAL	14	14	14	25	0	0
Campo 41	DDDP_DDD_NUMERO_DESTINO	61	64	64	64	64	64
Campo 42	DDDP_PREF_NUMERO_DESTINO	9973	9957	9952	8424	3613	3621
Campo 43	DMOD_CODIGO	M	M	M	M	K	J
Campo 44	SENTIDO_CHAMADA	E	S	S	S	E	E
Campo 45	DERE_CODIGO						
Campo 46	DTAR_CODIGO		N	N		N	N
Campo 47	CONC_PESS_ID_ORIGEM	4323	2715	2715	2000182	641123	2000388
Campo 48	CONC_PESS_ID_DESTINO	4323	4215	4215	2000492	2715	2715
Campo 49	CABI_CODIGO	13783	8798	8798	106491	143819	159054
Campo 50	CSRE_CODIGO	204225	142670	142670	200895	164769	59597
Campo 51	CTAR_CODIGO	2547					
Campo 52	CODIGO_OPERADORA						
Campo 53	BILHETADOR						
Campo 54	ANUF_POI						
Campo 55	CATEGORIA_DESTINO						
Campo 56	INTE_NR_INTERVALO_ORIGEM	1	1	1	3	1	1
Campo 57	INTE_NR_INTERVALO_DESTINO	1	1	1	1	1	1
Campo 58	TCTE_CODIGO_DESTINO	10	10	10	10	10	10
Campo 59	TUTE_CODIGO_ORIGEM	0	43	43	0	0	0
Campo 60	TUTE_CODIGO_DESTINO	0	0	0	0	0	0
Campo 61	CODIGO_BILHETADORA						
Campo 62	LOCALIDADE_ORIGEM						
Campo 63	LOCALIDADE_DESTINO						
Campo 64	OPERADORA_ORIGEM						
Campo 65	OPERADORA_DESTINO						
Campo 66	NUMERO_REDE_ORIGEM						
Campo 67	NUMERO_REDE_DESTINO						
Campo 68	DATA_HORA_COBRANCA	01Nov2009	01Nov2009	01Nov2009	01Nov2009	01Nov2009	01Nov2009
Campo 69	CSRE_CODIGO_2						

3.2. DIMENSIONAMENTO DA AMOSTRA

A amostra utilizada para demonstrar o processo é constituída pelo tráfego de Longa Distância gerado por uma cidade com mais de 1 milhão de habitantes atendidos por uma rede telefônica com 42 centrais de comutação[13] onde cada arquivo de tráfego diário tem em média 8 milhões de CDR. O primeiro arquivo de CDR completo exportado para o SAS ocupou 1,3 Gb de memória somente como input e à medida que o arquivo foi sendo processado mais 350 Mb foi consumida. A estimativa para processar um mês de tráfego foi de 52 Gb e para os quatro meses de tráfego que obtidos em Costa (2010), necessitaria 208 Gb o que tornaria inviável o processamento desses dados em notebook de um analista de tráfego. Para viabilizar o trabalho em um ambiente de análise tráfego das operadoras, foram realizadas algumas reduções. Inicialmente foram analisados todos os campos do CDR e selecionados aqueles necessários e suficientes para garantir o processo de análise e projeção do tráfego nos níveis de filial, central e hora da chamada. Além desta redução, simplificou-se o cam

segundo e durante o processo de carga para o SAS, apenas a informação da hora cheia foi transportada. Exemplificando, todas as chamadas de uma determinada central originadas entre 8:00:00 (oito horas, zero minutos e zero segundos) e 8:59:59 (oito horas, cinqüenta e nove minutos e cinqüenta e nove segundos) foram carregadas no SAS como realizadas na hora 8. A amostra escolhida ficou apenas 14 campos como apresentada tabela 3.2.

Tabela 3.2 Amostra dos dados disponíveis de tráfego diário

CAMPO 1	CAMPO 2	CAMPO 3	CAMPO 4	CAMPO 5	CAMPO 6	CAMPO 7	CAMPO 8	CAMPO 9	CAMPO 10	CAMPO 11	CAMPO 12	CAMPO 13	CAMPO 14
CSRE_CODIGO	CTAR_CODIGO	CON_PES_ID_ORIG	CON_PES_ID_DEST	CENTRAL	DIA_MES	TIPO_TRAFEGO	PARTE_TARIF	COD_SEL_NAC	HORA	REM	TAR	CBIL	DURACAO_REAL
1	0	4328	2715	CD2	31/01/10	4	0	0	11	1	0	1	0,47
1	0	4328	2715	CD2	31/01/10	6	0	21	8	1	0	1	1,83
1	0	4328	2715	CD2	31/01/10	6	0	21	10	1	0	1	0,67
1	0	4328	2715	CD2	31/01/10	6	0	21	9	1	0	1	0,92

A redução da amostra de 69 para 14 campos foi realizada fora do SAS e com , sendo necessário ler cada um dos 120 arquivos diários e selecionar apenas os 14 campos escolhidos. O trabalho de redução foi realizado com a colaboração de Costa (2010) e com ajuda de especialistas em TI porque foram necessários conhecimentos específicos de processamento dados e capacidade de máquina maior que os notebook utilizados por analistas de tráfego, além de não fazer parte do escopo do processo de análise e projeção.

[13] Central de comutação: elemento que interliga os usuários com a rede telefônica e gera um CDR para cada ligação originada por um usuário.

Carregando a amostra menor, com apenas 14 campos, o espaço de memória necessário foi reduzido para 300 Mb por dia de tráfego. Á medida que o arquivo com esta amostra foi sendo processado pelo SAS, mais memória foi sendo consumida, de modo que após as diversas etapas de configuração, seleção, agrupamento, análise e projeção mais 350 Mb foi necessário. Em resumo, para processar um mês de tráfego com input de 14 campos foi necessário 20 Gb[14] de memória do notebook utilizando na dissertação foi possível carregar o SAS com tráfego referente a dois meses ou 59 dias de tráfego, referente ao período de 1 de janeiro de 2010 a 28 de fevereiro de 2010, com a utilização de 40 Gb de memória.

Apesar do grande volume de dados, uma vez tendo carregado o tráfego no SAS, o processamento, das informações foi muito eficiente. As tabelas de tráfego com até 7 milhões de registros de 14 campos cada, foram processadas pelo SAS sem dificuldades e de uma forma muito rápida. A versão utilizada do SAS Enterprise Guide 4.2 é uma aplicação voltada para o cliente Microsoft Windows que dar possibilidade ao usuário para usar todos os recursos do software sem saber programar em código SAS. Esta é uma das grandes vantagens da utilização desta versão do SAS, ou seja, profissionais das áreas de gerenciamento de tráfego, planejamento e faturamento podem utilizar o SAS como se estivessem utilizando o Office.

Após algumas simulações, reflexões e conversas com profissionais que trabalham com análise de tráfego, decidi apresentar um o processo de análise em quatro etapas: dia do mês; dia da semana; intervalo de hora do dia e central. A idéia é elaborar um processo para viabilizar a identificação de uma tendência no tráfego diário de uma cidade, de uma determinada filial, podendo descer a detalhes de dia da semana, intervalo de hora e central com agilidade, eficiência e eficácia.

Continuando o processo de dimensionamento da amostra, a próxima etapa é de seleção dos campos: dia do mês; dia da semana; intervalo de hora do dia e central para dia da amostra. A seleção feita no SAS é apresentada na tabela 3.3:

Tabela 3.3 Amostra após o processo de seleção de campos

CTAR_CODIGO	CENTRAL	DIA_MES	HORA	TAR	DURACAO_REAL
0	CD2	31/01/10	11	0	0,47
0	CD2	31/01/10	8	0	1,83
0	CD2	31/01/10	10	0	0,67
0	CD2	31/01/10	9	0	0,92

[14] Para o arquivo de um dia de tráfego necessitamos de 650Mb em média, sendo 300Mb para o arquivo de entrada e 350Mb para os arquivos processados durante a análise e projeção. Para um mês, a demanda é (300 Mb+350Mb) *30 = 19.500 Mb = 20Gb

Nela o campo CENTRAL traz o exemplo do nome de central telefônica onde as chamadas telefônicas foram originadas. O campo DIA_MES contém a informação do dia, mês e ano de início da chamada. O campo HORA representa o intervalo de hora no qual a chamada teve inicio com 24 opções de intervalos. O campo DURACAO representa a duração de cada chamada em centésimos de minutos. Os campos CTAR_CODIGO e TAR são campos que sinalizam as chamadas que devem ser tarifadas e encaminhadas para o processo de faturamento ou emissão de boletos.

A etapa seguinte foi escolher somente as chamadas prontas para serem tarifadas. As chamadas que atendem esta condição são aquelas que apresentam os campos CTAR_CODIGO e TAR com valores maiores que zero. Para obter esta condição utilizamos a função de filtragem do SAS para obter todas chamadas com esta condição satisfeita. O resultado está na tabela 3.4 que mostra os campos que formam a base de dados para o processo de análise do tráfego.

Tabela 3.4 Amostra após a filtragem de campos

CENTRAL	DIA_MES	HORA	DURACAO_REAL
CD2	31/01/10	13	0,97
CD2	31/01/10	14	1,93
CD2	31/01/10	13	0,08
CD2	31/01/10	13	1,80

Uma etapa intermediária no processo foi a identificação do dia da semana para cada campo DIA_MES. Isto é muito importante porque existe uma tendência do tráfego por dia da semana. Por exemplo, o tráfego do sábado e domingo normalmente é menor do que o tráfego de um dia útil. O modo mais eficiente de identificação do dia da semana foi através da codificação elementar no programa no SAS realizada com o auxilio de profissional de TI. A tabela 3.5 apresenta o formato da base de dados na versão final, preparada para o processo de análise do tráfego. Os campos: CENTRAL, HORA e DURACAO_REAL são os mesmos da tabela 3.4. Entre os novos, o o campo mostra o mês do tráfego, o campo o campo da mostra o dia da semana do tráfego.

Tabela 3.5 Amostra da base de dados com dia da semana

CENTRAL	HORA	DURACAO_REAL	dia	mes	data	dia_da_semana
CD2	13	0,97	31	1	31/01/2010	Domingo
CD2	14	1,93	31	1	31/01/2010	Domingo
CD2	13	0,08	31	1	31/01/2010	Domingo
CD2	13	1,8	31	1	31/01/2010	Domingo

Este formato é a estrutura final da base de dados que utilizada no processo de análise do tráfego.

3.3. PROCESSO DE PREPARAÇÃO DA BASE

Em resumo, a figura 3.2 apresenta a visão geral do processo de preparação da base

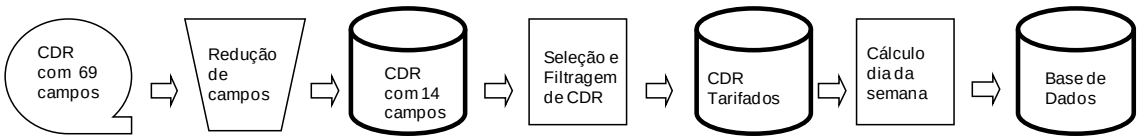


Figura 3.2 Processo de preparação da base de dados.

de dados a ser utilizada no processo de análise desde a o recebimento de arquivos coletados em Costa (2011) e utilizados no seu trabalho de dissertação até a obtenção de uma base de dados com as chamadas tarifadas. É destaque o fato que todo o processo foi executado pelo SAS instalado em notebook, exceto a etapa de redução dos 69 campos, que foi executada fora do SAS por falta de memória no HD. O fato de maior parte de o processo ter sido executada no ambiente do SAS, permite que a base de dados possa ser revisada para acrescentar novos campos com agilidade e autonomia pela equipe responsável pela análise e projeção do tráfego.

3.4. MODELOS DE ARQUIVOS DE ANÁLISE

Uma condição necessária para garantir que o processo de análise do tráfego tenha agilidade suficiente para explicar os desvios e distorções mensais da receita de voz, é flexibilidade que as ferramentas utilizadas têm em gerar relatórios diversificados. Um relatório importante é o que apresenta a distribuição do tráfego por hora do dia. Como exemplo, a tabela 3.6 mostra a distribuição do tráfego em minutos do dia 16/12/2009 para as 41 centrais de comutação da localidade escolhida como amostra do

Tabela 3.6 Distribuição do tráfego por intervalo de hora

data	dia_da_semana	HORA	SUM_of_DURACAO_REAL
16/12/2009	Quarta	0	134.770,80
16/12/2009	Quarta	1	31.883,16
16/12/2009	Quarta	2	33.801,04
16/12/2009	Quarta	3	26.900,04
16/12/2009	Quarta	4	16.957,38
16/12/2009	Quarta	5	39.116,69
16/12/2009	Quarta	6	36.906,10
16/12/2009	Quarta	7	94.065,65
16/12/2009	Quarta	8	404.388,62
16/12/2009	Quarta	9	760.401,13
16/12/2009	Quarta	10	727.908,49
16/12/2009	Quarta	11	566.599,76
16/12/2009	Quarta	12	420.100,37
16/12/2009	Quarta	13	459.624,65
16/12/2009	Quarta	14	556.351,41
16/12/2009	Quarta	15	568.265,95
16/12/2009	Quarta	16	572.872,83
16/12/2009	Quarta	17	511.134,13
16/12/2009	Quarta	18	337.446,21
16/12/2009	Quarta	19	299.608,93
16/12/2009	Quarta	20	329.249,15
16/12/2009	Quarta	21	330.896,52
16/12/2009	Quarta	22	228.977,16
16/12/2009	Quarta	23	157.397,08
Total do Dia			7.645.623,25

O tráfego total do dia é de, aproximadamente, 7,6 milhões de minutos para uma cidade com um pouco mais 1 milhão de habitantes, resultando em uma média de quase 6 minutos de tráfego diário -se uma variação do tráfego para cada intervalo de hora. Esta tabela foi obtida a partir da base apresentada na tabela 3.5 e utilizando a função query builder do SAS para selecionar os campos: data;

dia_da_semana; HORA e DURACAO_REAL. No campo DURACAO_REAL foi utilizada a função SUM para obter o somatório do tráfego para cada intervalo de hora. A tabela do SAS foi exportada para o Excel onde foi desenhada a forma final da tabela 3.6

Outra visualização importante dos dados é na forma gráfica que permite entender a variação diária do tráfego. Na figura 3.3 é apresentado um gráfico da distribuição do tráfego em minutos para cada intervalo de hora do dia 16/dez/2009 com as mesmas informações da tabela 3.6.

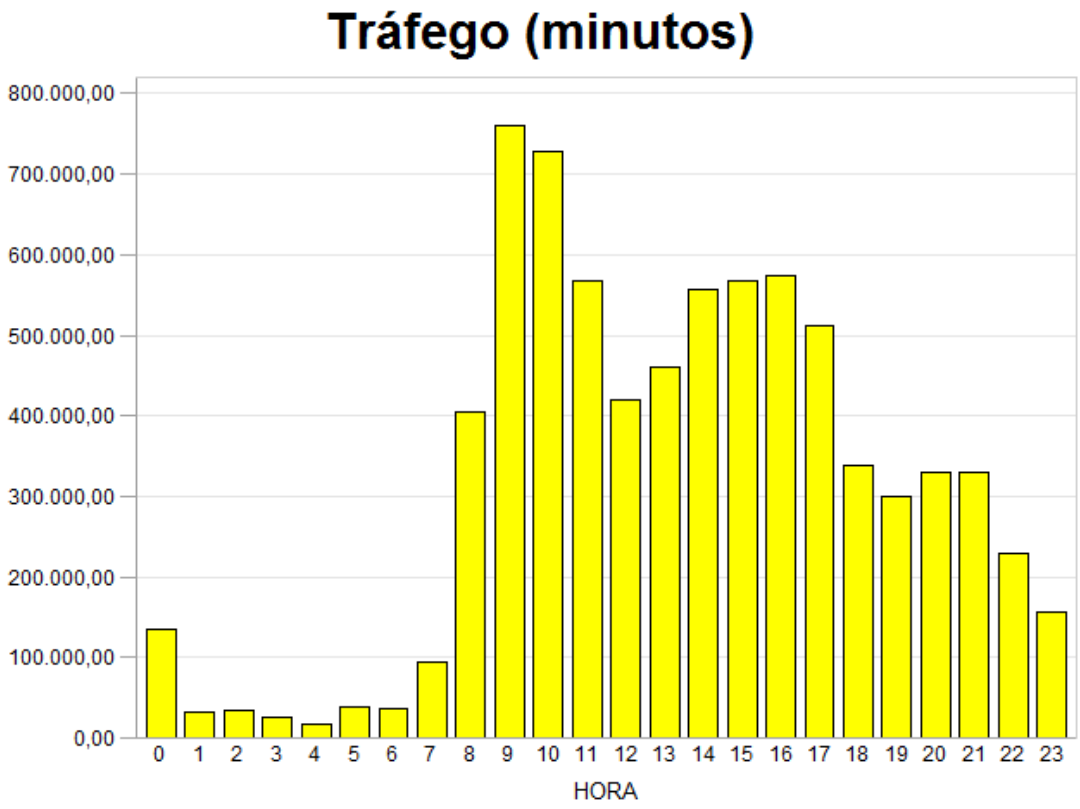


Figura 3.3 SAS: Distribuição do Tráfego por intervalo de hora.

O gráfico da figura 3.3 foi desenhado no SAS, utilizando a função BAR CHAT nos dados da tabela 3.6. Nesse gráfico, o tráfego do intervalo entre 0 e 7 horas o tráfego é desprezível enquanto que nos intervalos entre 9 e 11 horas e entre 14 e 16H é o mais relevante. A hora de maior movimento - H.M.M., o pico de tráfego, está localizada no intervalo de 9 horas.

Outra informação importante para análise do tráfego é a sua distribuição por central telefônica. Na tabela 3.7 é apresentada a informação do tráfego cursado em cada central e foi obtida a partir da base apresentada na tabela 3.5 e utilizando a função query builder do SAS para selecionar os campos: CENTRAL; data; dia_da_semana; e DURACAO_REAL.

No campo DURACAO_REAL foi utilizada a função SUM para obter o somatório do tráfego para cada central.

Tabela 3.7 Distribuição do tráfego por central

CENTRAL	data	dia_da_semana	SUM_of_DURACAO_REAL
AB1	16/12/2009	Quarta	139.586,05
AB2	16/12/2009	Quarta	224.510,40
AB3	16/12/2009	Quarta	91.798,35
AB4	16/12/2009	Quarta	45.637,70
AB5	16/12/2009	Quarta	197.142,53
AB6	16/12/2009	Quarta	130.944,14
AB7	16/12/2009	Quarta	297.293,65
AB8	16/12/2009	Quarta	47.524,19
AB9	16/12/2009	Quarta	648.467,17
CD1	16/12/2009	Quarta	156.210,69
CD2	16/12/2009	Quarta	157.207,74
CD3	16/12/2009	Quarta	440.360,53
CD4	16/12/2009	Quarta	72.914,85
CD5	16/12/2009	Quarta	450.410,34
CD6	16/12/2009	Quarta	2.909,64
CD7	16/12/2009	Quarta	43.121,88
CD8	16/12/2009	Quarta	185.890,90
CD9	16/12/2009	Quarta	190.619,15
EF1	16/12/2009	Quarta	24.572,20
EF2	16/12/2009	Quarta	100.850,94
EF3	16/12/2009	Quarta	692.019,47
EF4	16/12/2009	Quarta	279.781,31
EF5	16/12/2009	Quarta	34.162,51
EF6	16/12/2009	Quarta	158.738,14
EF7	16/12/2009	Quarta	145.545,30
EF8	16/12/2009	Quarta	182.628,47
EF9	16/12/2009	Quarta	142.407,28
GH1	16/12/2009	Quarta	198.087,84
GH2	16/12/2009	Quarta	24.706,45
GH3	16/12/2009	Quarta	549.525,70
GH4	16/12/2009	Quarta	158.155,35
GH5	16/12/2009	Quarta	169.466,77
GH6	16/12/2009	Quarta	15.234,86
GH7	16/12/2009	Quarta	138.880,89
GH8	16/12/2009	Quarta	306.318,23
GH9	16/12/2009	Quarta	370.118,29
IJ1	16/12/2009	Quarta	258.996,56
IJ2	16/12/2009	Quarta	62.963,69
IJ3	16/12/2009	Quarta	44.093,50
IJ4	16/12/2009	Quarta	40.567,48
IJ5	16/12/2009	Quarta	25.252,12
Total Dia			7.645.623,25

Observa-se uma grande variação no tráfego de cada central da tabela 3.8 e para analisar estas variações e observar quais são os maiores e menores valores, utiliza-se a tabela 3.9 que apresenta a classificação das centrais pela quantidade de tráfego.

Tabela 3.8 - Classificação das centrais por tráfego diário

CENTRAL	data	dia_da_semana	SUM_of_DURACAO_REAL
EF3	16/12/2009	Quarta	692.019,47
AB9	16/12/2009	Quarta	648.467,17
GH3	16/12/2009	Quarta	549.525,70
CD5	16/12/2009	Quarta	450.410,34
CD3	16/12/2009	Quarta	440.360,53
GH9	16/12/2009	Quarta	370.118,29
GH8	16/12/2009	Quarta	306.318,23
AB7	16/12/2009	Quarta	297.293,65
EF4	16/12/2009	Quarta	279.781,31
IJ1	16/12/2009	Quarta	258.996,56
AB2	16/12/2009	Quarta	224.510,40
GH1	16/12/2009	Quarta	198.087,84
AB5	16/12/2009	Quarta	197.142,53
CD9	16/12/2009	Quarta	190.619,15
CD8	16/12/2009	Quarta	185.890,90
EF8	16/12/2009	Quarta	182.628,47
GH5	16/12/2009	Quarta	169.466,77
EF6	16/12/2009	Quarta	158.738,14
GH4	16/12/2009	Quarta	158.155,35
CD2	16/12/2009	Quarta	157.207,74
CD1	16/12/2009	Quarta	156.210,69
EF7	16/12/2009	Quarta	145.545,30
EF9	16/12/2009	Quarta	142.407,28
AB1	16/12/2009	Quarta	139.586,05
GH7	16/12/2009	Quarta	138.880,89
AB6	16/12/2009	Quarta	130.944,14
EF2	16/12/2009	Quarta	100.850,94
AB3	16/12/2009	Quarta	91.798,35
CD4	16/12/2009	Quarta	72.914,85
IJ2	16/12/2009	Quarta	62.963,69
AB8	16/12/2009	Quarta	47.524,19
AB4	16/12/2009	Quarta	45.637,70
IJ3	16/12/2009	Quarta	44.093,50
CD7	16/12/2009	Quarta	43.121,88
IJ4	16/12/2009	Quarta	40.567,48
EF5	16/12/2009	Quarta	34.162,51
IJ5	16/12/2009	Quarta	25.252,12
GH2	16/12/2009	Quarta	24.706,45
EF1	16/12/2009	Quarta	24.572,20
GH6	16/12/2009	Quarta	15.234,86
CD6	16/12/2009	Quarta	2.909,64

Nessa tabela, as centrais EF3 e AB9 tem mais de 600.000 minutos de tráfego dia enquanto as centrais GH6 e CD6 tem menos de 20.000 minutos dia, ou seja uma diferença de 20 vezes. A tabela 3.8 foi obtida a partir da aplicação da função Filter and Sort na tabela anterior. Um formato mais geral e muito necessário é o tráfego para cada data da amostra onde é possível observar comportamento diário de todas as centrais da localidade escolhida como amostra, conforme apresentado na tabela 3.9 e na figura 3.4.

Tabela 3.9 Distribuição de tráfego por dia de semana

data	dia_da_semana	SUM_of_SUM_of_DURACAO_REAL
01/01/2010	Sexta	4.066.796,74
02/01/2010	Sábado	3.898.202,44
03/01/2010	Domingo	3.856.674,87
04/01/2010	Segunda	8.178.279,87
05/01/2010	Terça	8.518.410,89
06/01/2010	Quarta	8.296.551,71
07/01/2010	Quinta	7.958.836,52
08/01/2010	Sexta	8.006.706,62
09/01/2010	Sábado	5.089.719,53
10/01/2010	Domingo	4.864.757,20
11/01/2010	Segunda	8.277.379,69
12/01/2010	Terça	8.079.057,20
13/01/2010	Quarta	7.803.069,06
14/01/2010	Quinta	8.086.862,67
15/01/2010	Sexta	7.775.869,98
16/01/2010	Sábado	5.275.600,12
17/01/2010	Domingo	4.267.443,33
18/01/2010	Segunda	8.181.252,94
19/01/2010	Terça	7.855.737,56
20/01/2010	Quarta	7.549.287,16
21/01/2010	Quinta	7.924.079,60
22/01/2010	Sexta	7.299.360,03
23/01/2010	Sábado	4.812.510,26
24/01/2010	Domingo	3.804.233,25
25/01/2010	Segunda	8.201.779,52
26/01/2010	Terça	8.911.600,53
27/01/2010	Quarta	8.726.802,53
28/01/2010	Quinta	5.587.898,39
29/01/2010	Sexta	4.846.666,96
30/01/2010	Sábado	3.268.318,63
31/01/2010	Domingo	1.053.701,52
01/02/2010	Segunda	8.804.484,35
02/02/2010	Terça	8.893.631,09
03/02/2010	Quarta	8.938.565,57
04/02/2010	Quinta	8.862.706,93
05/02/2010	Sexta	8.846.371,34
06/02/2010	Sábado	6.454.431,35
07/02/2010	Domingo	5.782.710,40
08/02/2010	Segunda	9.264.814,64
09/02/2010	Terça	9.215.338,86
10/02/2010	Quarta	9.103.380,37
11/02/2010	Quinta	8.737.768,82
12/02/2010	Sexta	8.393.857,31
13/02/2010	Sábado	5.625.262,75
14/02/2010	Domingo	4.278.862,67
15/02/2010	Segunda	4.092.866,57
16/02/2010	Terça	4.143.948,03
17/02/2010	Quarta	7.279.958,07
18/02/2010	Quinta	8.393.862,23
19/02/2010	Sexta	8.141.334,43
20/02/2010	Sábado	5.711.056,61
21/02/2010	Domingo	4.733.249,95
22/02/2010	Segunda	8.414.801,49
23/02/2010	Terça	8.235.181,45
24/02/2010	Quarta	7.696.368,04
25/02/2010	Quinta	8.493.869,76
26/02/2010	Sexta	8.199.438,51
27/02/2010	Sábado	1.894.747,17
28/02/2010	Domingo	3.262.861,59

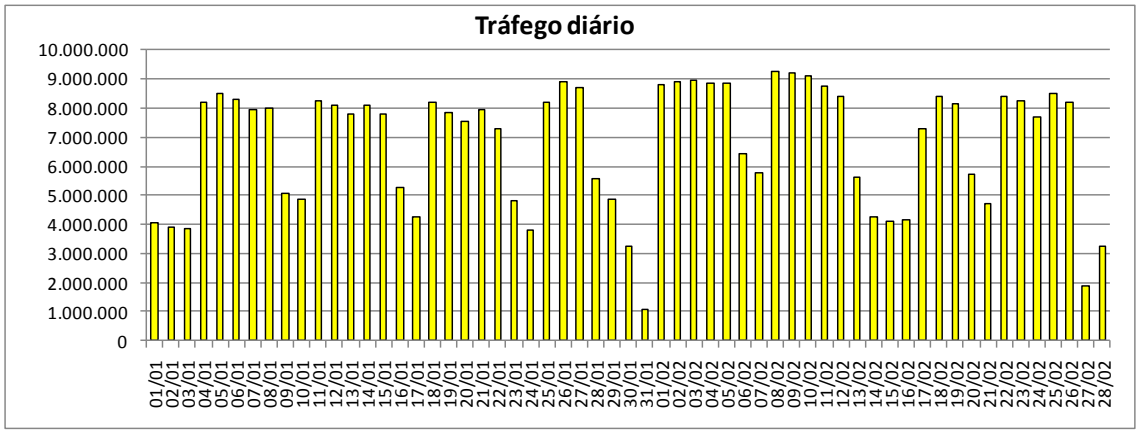


Figura 3.4 Distribuição do Tráfego por dia do mês.

A tabela 3.9 foi obtida a partir da base apresentada na tabela 3.5 e utilizando a função query builder do SAS para selecionar os campos: data; dia_da_semana; e DURACAO_REAL. No campo DURACAO_REAL foi utilizada a função SUM para obter o somatório do tráfego para cada dia do mês. O gráfico da figura 3.4 foi obtido utilizando a função gráfico de barras do Excel nos dados da tabela 3.9.

Os modelos apresentados acima são apenas alguns exemplos de tabelas e gráficos de análise que o SAS pode oferecer diretamente e em colaboração com o Excel. Uma equipe de analistas de tráfego com algum tempo de experiência, pode muito da capacidade do SAS para trabalhar com bastante eficiência e muito resultados

4. ANÁLISE DE TRÁFEGO

Neste capítulo descreve-se uma proposta de um processo para análise do tráfego telefônico desde a sua dimensão maior que é o tráfego da empresa, passando pelas filiais, cidades, centrais, dia e intervalo de hora. Aqui se apresentam as ferramentas utilizadas para análise dos dados e demonstração do resultado: SAS e Excel. O SAS com todo o seu pacote estatístico foi utilizado para fazer toda análise e pesquisa, enquanto o Excel serviu para configurar algumas tabelas e gráficos.

4.1. DIMENSÕES DA ANÁLISE DO TRÁFEGO

O processo de análise do tráfego tarifado[15], gerador da receita de voz das operadoras de Telecom, como exemplificado na tabela 1.1, é bastante complexo por envolver várias dimensões e um grande volume de dados. Este tráfego é caracterizado por uma chamada telefônica completada ou uma ligação atendida. Não se consideram neste estudo as tentativas de chamadas, ou seja, aquelas ligações não atendidas ou que encontraram a linha ocupada no destino. O tráfego é medido em minutos, e para entendê-lo e analisá-lo é necessário identificar todas estas dimensões e selecionar aquelas de maior impacto. A importância de analisar este tráfego está na necessidade de acompanhar o resultado financeiro mensal de uma Operadora de Telecom que tem na receita de voz uma parcela relevante da receita operacional e por isso necessita um processo para monitorar o tráfego em todas suas dimensões e acompanhar o processo de faturamento e contabilização dos CDR.

O tráfego é um reflexo direto e imediato do comportamento social dos indivíduos na sua necessidade de comunicação ampla e da atividade econômica dos indivíduos e das empresas. Uma dimensão importante do tráfego é o tempo. O tráfego varia 24 horas do dia e sete dias na semana. Dentro de um mesmo dia, o comportamento do tráfego oscila bastante, tendo seus valores máximos, nas HMM e valores mínimos durante a madrugada. As HMM podem variar entre cidades devido a atividade econômica de cada cidade e à diferença de fuso horário. Durante a semana, o tráfego pode variar com os dias sendo que a média diária de segunda a sexta é bem maior que média de sábados, domingos e feriados.

[15] Tráfego tarifado é tráfego gerado por chamadas que foram completadas e o CDR gerado será utilizado para a elaboração da fatura

Entretanto, em alguns feriados como dia das mães e Natal, observam-se as maiores médias do ano. Em um mesmo mês, o tráfego médio das semanas é diferente e os meses com mais dias apresentam um tráfego maior. Como exemplo, o tráfego do mês de fevereiro é em média 10% menor[16] que o tráfego de janeiro porque tem três dias a menos. A equipe que gerencia a receita de voz das Operadoras de Telecom convive com o fato de que a receita de voz de fevereiro é aproximadamente 10% menor que a de janeiro. Eventos incomuns como jogos pan-americanos, olimpíadas, campeonatos, feiras ou mesmo grandes acidentes produzem impactos relevantes no tráfego diário e mensal.

Outra dimensão do tráfego é a abrangência que pode ser local, regional, de longa distância ou internacional. O tráfego local tem as suas próprias dimensões que são de destino fixo ou móvel. O preço do minuto para uma chamada com destino fixo é menor que uma com destino móvel. O tráfego local para destino fixo é tarifado por multimedição.(Anatel, 1998). Neste processo, as chamadas telefônicas são tarifadas por pulsos. Um pulso, obrigatoriamente ocorre, quando a chamada é iniciada ou terminada. Outros pulsos ocorrem numa cadência de 240s enquanto durar a chamada. Em 2007, conforme previsto em Anatel (2006) a tarifação para o tráfego local mudou para o processo de bilhetagem onde para cada chamada são gerados os CDR que contêm entre informações, o tempo de duração da chamada.

O tráfego regional e o de longa distância têm as suas dimensões nos degraus tarifários que variam com a distância entre a origem e o destino da chamada. Quanto maior for o degrau, maior é o preço do minuto de chamada. O tráfego de longa distância tem ainda a sua dimensão da modulação horária com quatro tarifas diferenciadas pela hora do dia e pelo dia da semana. O tráfego internacional tem as dimensões do degrau e do transporte. O minuto da chamada para um mesmo destino varia com a empresa que transporta o tráfego.

No Brasil, as grandes operadoras de Telecom estão presentes em quase todas as regiões e operam com várias filiais que são formadas por diferentes estados brasileiros ou por grupo deles. Cada estado é formado por localidades de variadas culturas e têm empresas com atividades econômicas bastante diversificadas. Uma dimensão importante para análise do tráfego de Voz é a origem da chamada, ou seja, de qual filial ou de qual localidade o tráfego é originado. As diversidades são tantas que para uma análise completa

[16] O mês de Janeiro tem 31 dias e o de Fevereiro tem, normalmente 28 dias. O desvio entre os meses é de 3/31= 0,097 , significando que Fevereiro tem 10% menos dias.

do tráfego deve-se considerar todas as dimensões existentes, o que aumenta bastante a complexidade do processo.

Outro fator importante na análise é o volume de tráfego que se deve considerar em nossa análise, uma vez que envolve bilhões chamadas ou bilhões de minutos no mês para grandes operadoras como Telefonica/Vivo, TIM/Intelig, Embratel/Claro e Oi somente no Brasil. Este volume de dados demanda uma infra-estrutura e ferramentas adequadas para elaboração desta análise com já discutida no capítulo 3.

O processo de análise de tráfego proposto considera estas dificuldades sem perder o foco da eficiência e eficácia. O processo começa com análise do tráfego em nível mais geral e à medida do necessário abre-se a análise para um nível mais específico conforme apresentado na figura 4.1.

1. Empresa
   1.1. Filial ou grupo de filial
      1.1.1. Localidade ou grupo de localidade
         1.1.1.1. Dia do mês
            1.1.1.1.1. Central
               1.1.1.1.1.1. Intervalo de Hora
         1.1.1.2. Dia da semana
            1.1.1.2.1. Central
               1.1.1.2.1.1. Intervalo de Hora

Figura 4.1 Hierarquia de análise do tráfego telefônico

O nível mais geral da análise é o tráfego total da empresa e abertura seguinte é por filial. Em um primeiro momento analisa-se o tráfego total da empresa comparando com uma referência que está na base de dados.

4.2. PROCESSO DE ANÁLISE DO TRÁFEGO

Análise do tráfego de Voz é uma parte da atividade de gerenciamento do resultado mensal de receita de voz. A outra parte é a monitoração do processo de faturamento e contabilização dos CDR. A informação do tráfego cursado é uma condição necessária,

mas não suficiente para garantir a receita mensal de filial. Como exemplo, se durante o processo de emissão de faturas de uma determinada filial ou cidade, existe uma programação para desconto relativo a uma campanha mercadológica mensal, o tráfego cursado é o esperado sem desvios, mas a receita deverá ser menor, uma vez que o preço médio aplicado foi menor. Quando as contas (faturas) de uma determinada localidade são emitidas com erro, um estorno dos valores financeiro é processado e neste mês a receita é descasada do tráfego cursado. O tráfego cursado é o esperado, mas a receita é menor.

O processo aqui proposto atende a demanda maior que é a análise do tráfego encaminhado e devido a complexidade e abrangência dos dados é executado no sentido top-down, do geral para o mais específico.

No caso da amostra selecionada referente a uma cidade com pouco mais que 1 milhão de habitantes e com 42 centrais telefônicas, foi carregado no SAS o tráfego relativo ao período de 2 meses que é apresentado na figura 4.2. O tráfego de outros meses não foi carregado, devido a limitações de capacidade explicadas no item 3. Entretanto, os dias disponibilizados permitem demonstrar a metodologia para análise de tráfego proposta.

4.2.1. Recursos dos SAS

Com base nos recursos disponibilizados pelo SAS, elaborou-se uma análise padrão para a amostra do tráfego diário apresentada na figura 4.2:

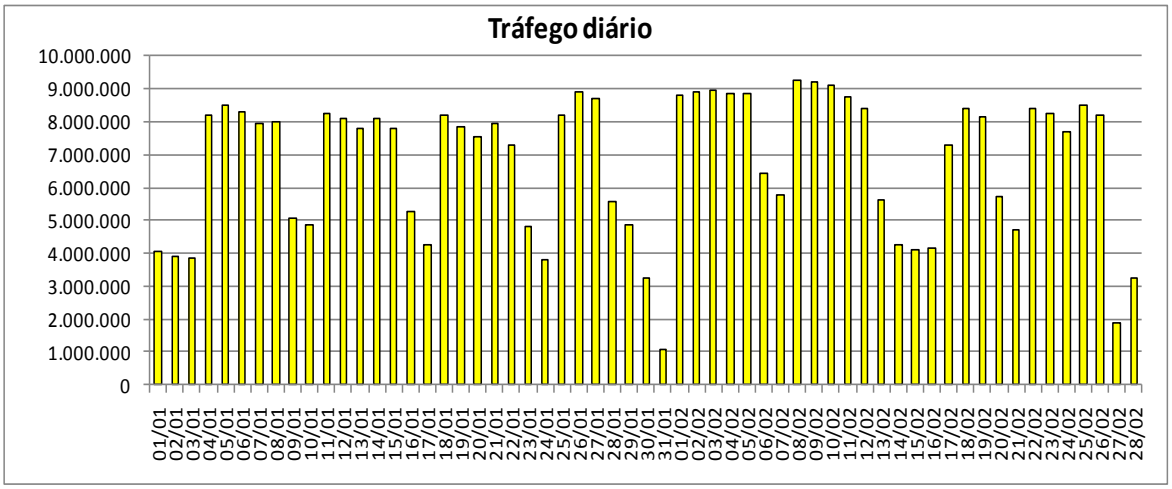


Figura 4.2 Distribuição do tráfego diário

O gráfico da figura 4.2 foi obtido a partir da base apresentada na tabela 3.5 e utilizando a função query builder do SAS para selecionar os campos: data; dia_da_semana; e DURACAO_REAL. No campo DURACAO_REAL foi utilizada a função SUM para obter o somatório do tráfego para cada dia do mês. Estas informações foram exportadas para o Excel, onde utilizando a função gráfico de Barras, obteve-se o gráfico do tráfego diário.

Como demonstração dos principais recursos de análise do SAS, aplicou-se a função Distribution Analysis às mesmas informações do tráfego diário da figura 4.2. O resultado é apresentado nas tabelas e figuras a seguir, obtidas diretamente do SAS. A tabela 4.1 apresenta alguns parâmetros básicos e entre eles, os Momentos de distribuição de probabilidade, que segundo Borror et al. (2006, p. 42), são parâmetros que descrevem as distribuições de probabilidade, medem suas propriedades e em certos casos especificam estas distribuições.

Tabela 4.1 Parâmetros estatísticos do SAS: Momentos

Moments			
N	59	**Sum Weights**	59
Mean	6783375.89	**Sum Observations**	400219178
Std Deviation	2122557.77	**Variance**	4.50525E12
Skewness	-0.8005511	**Kurtosis**	-0.4998987
Coeff Variation	31.2905816	**Std Error Mean**	276333.485

Destaca-se nessa tabela que indica a quantidade de eventos na amostra, no caso de, 59 dias. O parâmetro Sum Weights é a soma dos pesos de cada observação e neste caos a soma é 59 porque cada observação tem peso unitário.

média aritmética do tráfego diário, no valor de 6.783.375,89 minutos. O parâmetro Sum Observations representa a soma do tráfego de todas as observações, no caso 400.219.178 minutos. O parâmetro Variance é variância, indicador da medida da dispersão da amostra que no caso vale 4,5 x 10^{12} minutos-quadrados. A fórmula de cálculo da variância e dos outros parâmetros citados neste capítulo estão descritas no anexo D com base em Borror et al. (2006). [17] [18] indicam o desvio padrão, calculado em função da variância, com valor de 2.122.557,77 minutos e a relação

[17] Descrito no anexo D de acordo com Borror et al. (2006)
[18] Descrito no anexo D de acordo com Borror et al. (2006)

percentual do desvio com a média aritmética no valor de 31,29%, respectivamente. Os [19] [20] representam a estimativa de assimetria e curtose com valores de -0,80 e -0,50 respectivamente, onde valores negativos de assimetria caracterizam uma cauda assimétrica no sentido a esquerda da média. A curtose descreve a presença relativa de picos na distribuição de freqüências em comparação a uma distribuição normal. Os valores negativos são associados a distribuições achatadas e valores positivos significam presenças de picos na distribuição. Uma distribuição Normal tem curtose zero enquanto uma distribuição uniforme tem -1,2. O parâmetro Std Error Mean [21] é o erro médio padrão, calculado como desvio padrão, mas dividido pela raiz quadrada de N.

A tabela 4.2 apresenta os parâmetros básicos como Mean , Std Deviation e Variance , já citados e comentados e o Median [22] representando a mediana e indicando que metade das observações da amostra tem valores maiores que o seu e outra metade tem valores menores.

Tabela 4.2 Parâmetros estatísticos do SAS: medidas básicas

Basic Statistical Measures			
Location		Variability	
Mean	6783376	Std Deviation	2122558
Median	7855738	Variance	4.50525E12

observações com maiores valores e as cinco observações com menores obtidas da amostra apresentada na tabela 3.5. A observação com menor valor na amostra foi a 31 com 1.053.702 minutos de tráfego e a maior foi a 39 com o valor de 9.264.815 minutos de tráfego.

[19] Descrito no anexo D de acordo com Borror et al. (2006)
[20] Descrito no anexo D de acordo com Borror et al. (2006)
[21] Descrito no anexo D de acordo com Borror et al. (2006)
[22] Descrito no anexo D de acordo com Borror et al. (2006)

Tabela 4.3 Parâmetros estatísticos do SAS: observações extremas

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1053702	31	8911601	26
1894747	58	8938566	34
3262862	59	9103380	41
3268319	30	9215339	40
3804233	24	9264815	39

E s cinco maiores e os cinco menores valores da amostra apresentada na tabela 3.5. O menor valor da amostra foi de 1.053.702 minutos de tráfego e o maior foi de 9.264.815 minutos de tráfego.

Tabela 4.4 Parâmetros estatísticos do SAS: valores extremos

Extreme Values			
Lowest		Highest	
Order	Value	Order	Value
1	1053702	55	8911601
2	1894747	56	8938566
3	3262862	57	9103380
4	3268319	58	9215339
5	3804233	59	9264815

A diferença entre as tabelas 4.4 e 4.3 é que na de observações extremas podem aparecer, por exemplo, observações distintas com mesmo valor, enquanto na outra, aparecem somente os valores distintos.

Frequency Counts ndo as

de cada amostra, ou

percentual de cada contagem e nesta amostra como não existem observações com mesmo

valor, o percentual é sempre

.

Tabela 4.5 Parâmetros estatísticos do SAS: Contagem de frequências

Frequency Counts			
		Percents	
Value	Count	Cell	Cum
1053701.52	1	1.7	1.7
1894747.17	1	1.7	3.4
3262861.59	1	1.7	5.1
3268318.63	1	1.7	6.8
3804233.25	1	1.7	8.5
3856674.87	1	1.7	10.2
3898202.44	1	1.7	11.9
4066796.74	1	1.7	13.6
4092866.57	1	1.7	15.3
4143948.03	1	1.7	16.9
4267443.33	1	1.7	18.6
4278862.67	1	1.7	20.3
4733249.95	1	1.7	22.0
4812510.26	1	1.7	23.7
4846666.96	1	1.7	25.4
4864757.20	1	1.7	27.1
5089719.53	1	1.7	28.8
5275600.12	1	1.7	30.5
5587898.39	1	1.7	32.2
5625262.75	1	1.7	33.9
5711056.61	1	1.7	35.6
5782710.40	1	1.7	37.3
6454431.35	1	1.7	39.0
7279958.07	1	1.7	40.7
7299360.03	1	1.7	42.4
7549287.16	1	1.7	44.1
7696368.04	1	1.7	45.8
7775869.98	1	1.7	47.5
7803069.06	1	1.7	49.2
7855737.56	1	1.7	50.8
7924079.60	1	1.7	52.5
7958836.52	1	1.7	54.2
8006706.62	1	1.7	55.9
8079057.20	1	1.7	57.6
8086862.67	1	1.7	59.3
8141334.43	1	1.7	61.0
8178279.87	1	1.7	62.7
8181252.94	1	1.7	64.4
8199438.51	1	1.7	66.1
8201779.52	1	1.7	67.8

Frequency Counts			
		Percents	
Value	Count	Cell	Cum
8235181.45	1	1.7	69.5
8277379.69	1	1.7	71.2
8296551.71	1	1.7	72.9
8393857.31	1	1.7	74.6
8393862.23	1	1.7	76.3
8414801.49	1	1.7	78.0
8493869.76	1	1.7	79.7
8518410.89	1	1.7	81.4
8726802.53	1	1.7	83.1
8737768.82	1	1.7	84.7
8804484.35	1	1.7	86.4
8846371.34	1	1.7	88.1
8862706.93	1	1.7	89.8
8893631.09	1	1.7	91.5
8911600.53	1	1.7	93.2
8938565.57	1	1.7	94.9
9103380.37	1	1.7	96.6
9215338.86	1	1.7	98.3
9264814.64	1	1.7	100.0

A tabela 4.6 apresenta os quantis[23] das observações da amostra, onde estão relacionados os percentis e quartis da amostra apresentada na tabela 3.5. O 100º. percentil é o valor máximo da amostra de 9.264.815 minutos de tráfego. O 75º. percentil é o terceiro quartil no valor de 8.393.862 minutos de tráfego. O 50º. percentil é a mediana da amostra no valor de 7.855.738 minutos de tráfego. O 25º. percentil é o primeiro quartil no valor de 4.846.667 minutos de tráfego. O 0º. percentil é o valor mínimo da amostra de 1.053.702 minutos de tráfego.

Tabela 4.6 Parâmetros estatísticos do SAS: Quantis

Quantiles	
Quantile	Estimate
100% Max	9264815
99%	9264815
95%	9103380
90%	8893631
75% Q3	8393862
50% Median	7855738
25% Q1	4846667

[23] Descrito no anexo D de acordo com The Math Forum (2002)

Quantiles	
Quantile	**Estimate**
10%	3856675
5%	3262862
1%	1053702
0% Min	1053702

A tabela 4.7 apresenta o resultado de três testes para saber se a amostra representa uma população com distribuição normal. Segundo Schlotzhauer (2009, p.140) o mais recomendado é o teste de Shapiro-Wilk[24] e na ausência deste o Anderson-Darling, podendo ainda ser utilizado também o Kolmogorov-Smirnov.

Tabela 4.7 Parâmetros estatísticos do SAS:
Teste para Normalidade

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.869486	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.225774	**Pr > D**	<0.0100
Anderson-Darling	**A-Sq**	3.200346	**Pr > A-Sq**	<0.0050

Esta tabela apresenta o o tipo de teste para Normalidade realizado. Segundo Schlotzhauer (2009, p.140) o SAS somente utiliza o teste Shapiro-Milk para amostras com tamanho máximo de 2000 observações. O identifica a estatística de teste[25] e o seu correspondente valor. O principal parâmetro da tabela é o $\rho-$

probabilidade ($\rho-$value) podem variar de 0 a 1 ($0<\rho<1$).Valores muito próximos de zero indicam que a amostra não é proveniente de uma distribuição Normal. Neste caso o $\rho-$ de 0,0001 é muito próximo de zero, significando que a hipótese da amostra da tabela 3.5 representar uma população com distribuição Normal é muito fraca.

A figura 4.3 apresenta o Histograma do tráfego da amostra da tabela 3.5, onde os valores são automaticamente agrupados em intervalos de 1.500.000 de minutos de tráfego. O SAS apresenta também na mesma figura uma curva normal ajustada com base na média e o no desvio padrão da amostra da tabela 4.2.

[24] Descrito no anexo A de acordo com The Math Forum (2002)
[25] As estatísticas de teste estão definidas nos anexos A,B e C

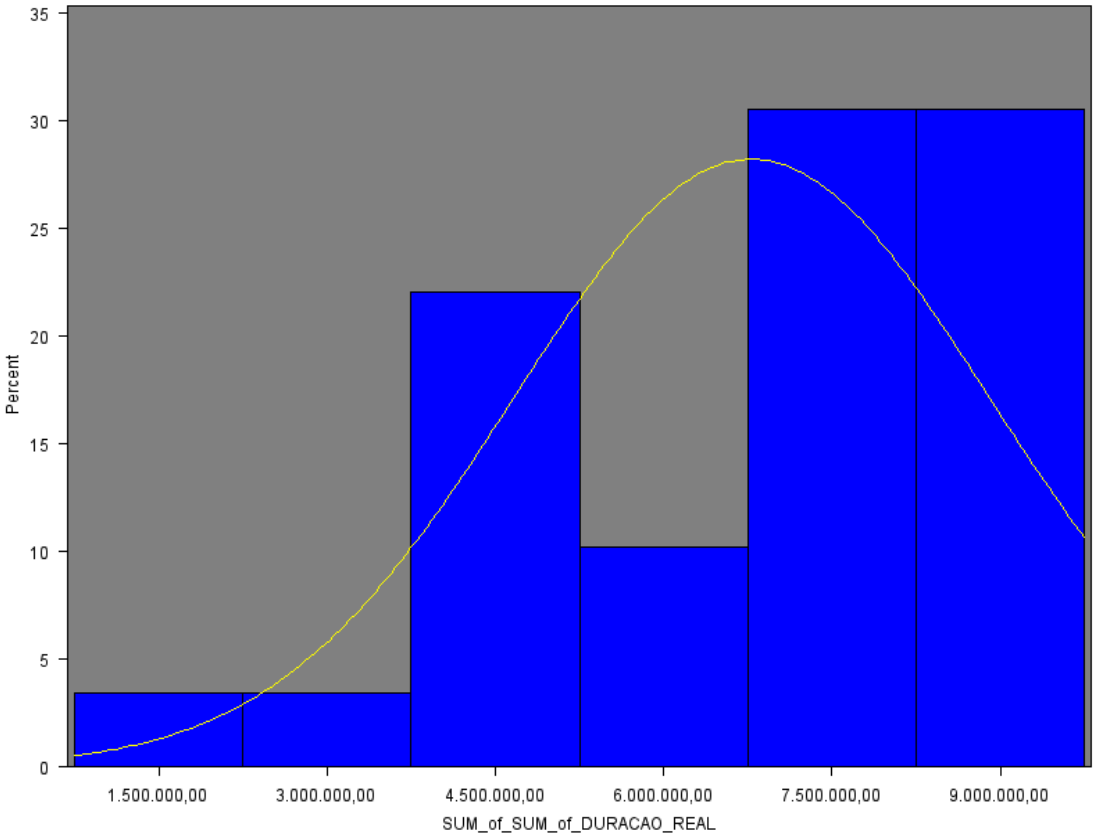


Figura 4.3 SAS: Histograma da amostra de 59 dias

O histograma da amostra não tem o formato do sino que caracteriza a distribuição Normal. Este resultado que é confirmado pelo teste de Shapiro-Wilk da tabela 4.7. Além disto, o histograma tem o formato assimétrico com uma cauda maior para esquerda e apresenta 3 picos relativos. Estes resultados são confirmados pelos parâmetros.

da tabela 4.1

A figura 4.4 apresenta o Diagrama Ramo-e-Folhas para o tráfego da amostra da tabela 3.5 numa escala de 10^6,

calculados pelo SAS que também fez a ordenação

folha tem a freqüência das observações correspondente ao intervalo escolhido. O SAS apresenta também na mesma figura um diagrama de Caixas que utiliza a escala do diagrama de Ramos-e-Folhas.

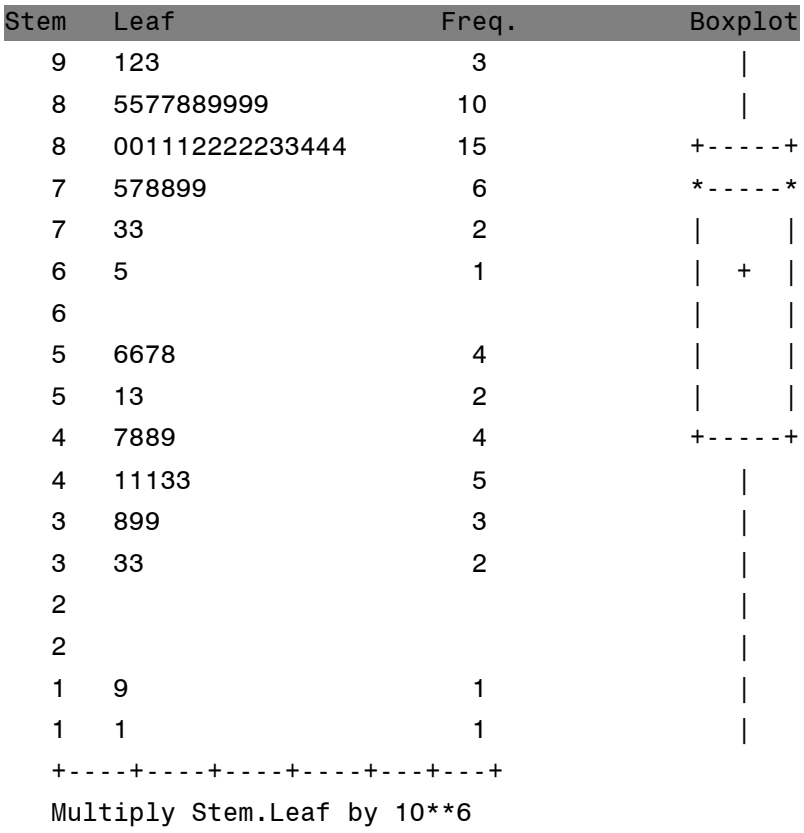


Figura 4.4 SAS: Diagrama de Ramos-e-Folhas para amostra de 59 dias

No diagrama de Caixas a linha tracejada inferior representa o primeiro quartil e a linha superior o terceiro quartil. A linha dentro da caixa com asterisco representa a mediana. A vantagem do diagrama de Ramos-e-Folha é os valores das amostras não são perdidos como no caso do histograma.

4.2.2. Identificando os pontos fora da curva

Analisando as figuras 4.3 e 4.4, percebe-se que os valores da amostra em análise estão distribuídos em três intervalos distintos. Esta distribuição é mais bem percebida quando todos os valores contidos em um mesmo intervalo são identificados com uma mesma graduação de cor, diferente da graduação de cores dos outros intervalos. A figura 4.5 apresenta a distribuição diária do tráfego com uma distinção dos valores pela graduação da cor.

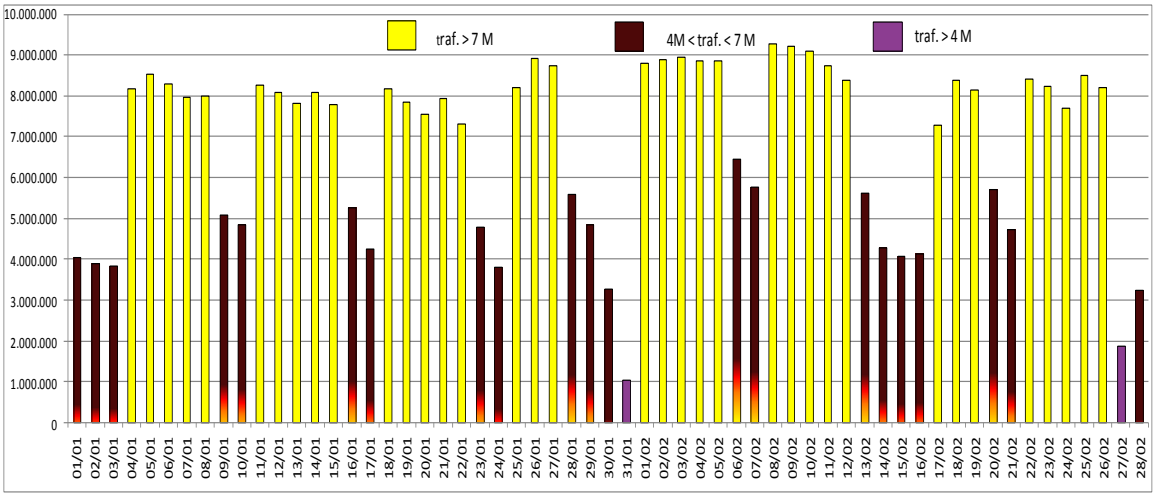


Figura 4.5 Distribuição de tráfego com marcação dos pontos fora da curva

Na cor sem graduação ou amarela, estão marcados os dias com valores de tráfego acima de 7 milhões de minutos . Em graduação mais forte na base ou vermelho estão os com tráfego menor que 7 milhões e maior que 3 milhões e com graduação mais fraca na base ou azul estão os valores menores que 3 milhões.

O próximo passo é dividir a amostra de acordo com os novos intervalos e avaliar as novas amostras com os recursos do SAS. Iniciando o processo, aplicou-se a função Distribution Analysis à amostra dos dias com tráfego diário maior que 7 milhões de minutos e o resultado obtido está apresentado a seguir. A tabela 4.8 apresenta a relação de todas observações que caracterizam esta nova amostra com dias de tráfego maior que 7 milhões de minutos.

Tabela 4.8 SAS: Contagem de freqüência para tráfego maior que 7 milhões

Frequency Counts			
		Percents	
Value	Count	Cell	Cum
7279958.07	1	2.8	2.8
7299360.03	1	2.8	5.6
7549287.16	1	2.8	8.3
7696368.04	1	2.8	11.1
7775869.98	1	2.8	13.9
7803069.06	1	2.8	16.7
7855737.56	1	2.8	19.4
7924079.60	1	2.8	22.2
7958836.52	1	2.8	25.0

Frequency Counts			
		Percents	
Value	Count	Cell	Cum
8006706.62	1	2.8	27.8
8079057.20	1	2.8	30.6
8086862.67	1	2.8	33.3
8141334.43	1	2.8	36.1
8178279.87	1	2.8	38.9
8181252.94	1	2.8	41.7
8199438.51	1	2.8	44.4
8201779.52	1	2.8	47.2
8235181.45	1	2.8	50.0
8277379.69	1	2.8	52.8
8296551.71	1	2.8	55.6
8393857.31	1	2.8	58.3
8393862.23	1	2.8	61.1
8414801.49	1	2.8	63.9
8493869.76	1	2.8	66.7
8518410.89	1	2.8	69.4
8726802.53	1	2.8	72.2
8737768.82	1	2.8	75.0
8804484.35	1	2.8	77.8
8846371.34	1	2.8	80.6
8862706.93	1	2.8	83.3
8893631.09	1	2.8	86.1
8911600.53	1	2.8	88.9
8938565.57	1	2.8	91.7
9103380.37	1	2.8	94.4
9215338.86	1	2.8	97.2
9264814.64	1	2.8	100.0

Essas novas observações formam uma nova amostra que será analisada e comparada com a amostra inicial apresentada na figura 4.2 e tabela 4.5

A tabela 4.9 apresenta os momentos de distribuição e ouros parâmetros estatísticos.

Tabela 4.9 SAS: Momentos para tráfego maior que 7 milhões

Moments			
N	36	**Sum Weights**	36
Mean	8320740.48	**Sum Observations**	299546657
Std Deviation	509837.463	**Variance**	2.59934E11
Skewness	-0.0550665	**Kurtosis**	-0.515516
Coeff Variation	6.1273088	**Std Error Mean**	84972.9106

Esta nova amostra, somente com tráfego maior que 7 milhões de minutos, é de 36 dias, com uma média de 8.320.740,48 minutos, o desvio padrão de 509.837,46 minutos e a relação percentual entre eles de 6,18%. Comparando com a amostra anterior de 59 dias, verifica-se que o desvio padrão desta nova amostra é bem menor, quase ¼, da amostra anterior de 2.122.557,77 minutos e que relação percentual de é apenas 6% contra quase 31%. É possível afirmar, segundo Borror et al. (2006, p.166) que os dados desta nova amostra tem uma variabilidade bem menor que a da anterior, com uma grande concentração de valores em torno da média. Seguindo com a análise verifica-se que a estimativa de Assimetria apresentou o valor -0,06, muito próximo de zero que segundo caracteriza uma simetria de curva Normal. Entretanto a estimativa de Curtose foi de -0,52 que indica uma distribuição achatada.

A tabela 4.10 apresenta a Mediana com o valor de 8.256.281 minutos de tráfego maior do que o valor 7.855.738 da amostra anterior apresentado na tabela 4.2 e bem próximo ao valor da Média de 8.320.740 com uma diferença de apenas 0,77%

Tabela 4.10 SAS: Parâmetros básicos para tráfego maior que 7 milhões

Basic Statistical Measures			
Location		**Variability**	
Mean	8320740	**Std Deviation**	509837
Median	8256281	**Variance**	2.59934E11

A tabela 4.11 apresenta as observações extremas relacionando as cinco observações com maiores valores que são iguais às cinco maiores observações da amostra anterior apresentadas na tabela 4.3. Entretanto, as cinco menores observações são maiores do que as da amostra anterior, como o esperado pela própria característica na nova amostra escolhida

Tabela 4.11 SAS: Observações extremas para tráfego maior que 7 milhões

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
7279958	29	8911601	17
7299360	15	8938566	21
7549287	13	9103380	26
7696368	34	9215339	25
7775870	10	9264815	24

A tabela 4.12 apresenta os valores extremos da amostra para o tráfego maior que 7 milhões de minutos relacionando os cinco maiores valores que são iguais aos cinco maiores observações da amostra anterior apresentadas na tabela 4.4. Entretanto, os cinco menores valores são maiores do que as da amostra anterior, como o esperado pela própria característica na nova amostra escolhida

Tabela 4.12 SAS: Observações extremas para tráfego maior que 7 milhões

Extreme Values			
Lowest		Highest	
Order	Value	Order	Value
1	7279958	32	8911601
2	7299360	33	8938566
3	7549287	34	9103380
4	7696368	35	9215339
5	7775870	36	9264815

A tabela 4.13 apresenta os quantis da nova amostra observações da amostra, onde estão relacionados os percentis de maior relevância e os quartis.

Tabela 4.13 SAS: Quantis para tráfego maior que 7 milhões de minutos

Quantiles	
Quantile	Estimate
100% Max	9264815
99%	9264815
95%	9215339
90%	8938566
75% Q3	8771127
50% Median	8256281
25% Q1	7982772
10%	7696368
5%	7299360
1%	7279958
0% Min	7279958

O 100º. percentil é o mesmo que o da amostra anterior no valor 9.264.815 minutos de tráfego. Entretanto, o 75º. percentil ou terceiro quartil tem o valor de 8.771.127 minutos e maior que o valor da amostra anterior de 8.393.862 apresentado na tabela 4.5. Assim também, o 50º. percentil ou mediana tem o valor de 8.256.281 minutos de tráfego maior que o valor de 7.855.738 da amostra anterior, o 25º. percentil ou primeiro quartil tem o valor de 7.982.772 minutos maior que o valor de 4.846.667 da amostra anterior e o 0º. percentil ou valor mínimo da amostra de 7.279.958 minutos de tráfego bem maior que o valor de 1.053.702 da amostra anterior.

A figura 4.6 apresenta o Histograma do tráfego da amostra da tabela 4.8, onde os valores são automaticamente agrupados em intervalos de 300.000 de minutos de tráfego.

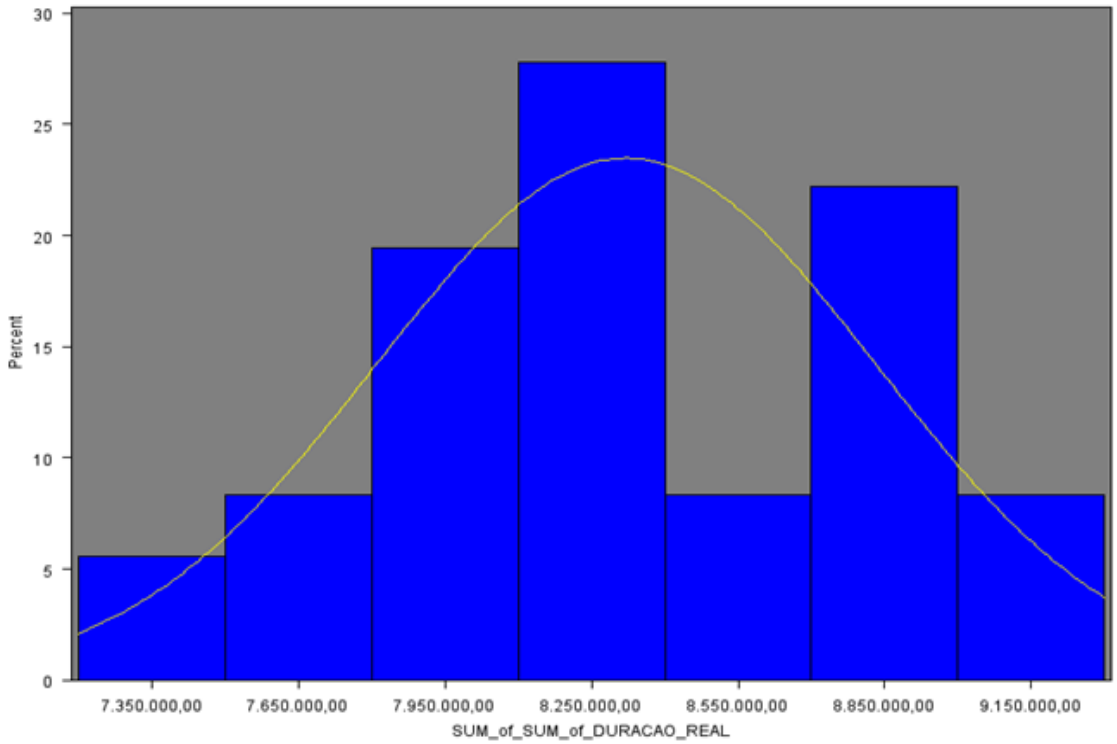


Figura 4.6 SAS: Histograma para tráfego maior que 7 milhões de minutos

O histograma desta amostra tem o formato mais próximo ao de um sino do que o da amostra anterior, apresentando uma simetria próxima a da distribuição Normal conforme indicado pelo parâmetro de Assimetria da tabela 4.9. Entretanto, existe uma diferença para o comportamento Normal que é existência de 2 picos relativos que confirma o resultado da

Curtose da tabela 4.9. Além disto, o SAS apresenta também na mesma figura uma curva normal ajustada com base na média e o no desvio padrão da tabela 4.9

Um resultado muito relevante é do teste para Normalidade de Shapiro-Wilk apresentado na tabela 4.14 com um ρ–Value de 0,6658 bem mais próximo de 1 do que ρ–Value da amostra anterior que foi de 0,0001, ou seja, a hipótese de uma distribuição Normal é mais provável para a amostra de 36 dias

Tabela 4.14 SAS: Teste para Normalidade para tráfego maior que 7 milhões de minutos

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.977656	**Pr < W**	0.6658
Kolmogorov-Smirnov	**D**	0.092671	**Pr > D**	>0.1500
Anderson-Darling	**A-Sq**	0.260784	**Pr > A-Sq**	>0.2500

A figura 4.7 apresenta o Diagrama Ramo-e-Folhas para o tráfego da amostra da tabela 4.8 numa escala de 10^5, onde a distribuição dos valores mostra uma concentração em torno do ramo $82x10^5$, mas ainda com existência de dois picos nos ramos $80x10^5$ e $88x10^{5,}$, conforme demonstrado pela Curtose da tabela 4.9 e mostrado no Histograma da figura 4.6.

```
Stem Leaf                    Freq.          Boxplot
  92 26                        2              |
  90 0                         1              |
  88 056914                    6              |
  86 34                        2            +-----+
  84 192                       3            |     |
  82 0048099                   7            *--+--*
  80 189488                    6            |     |
  78 0626                      4            +-----+
  76 08                        2              |
  74 5                         1              |
  72 80                        2              |
  +----+----+----+----+----+---+
  Multiply Stem.Leaf by 10**+5
```

Figura 4.7 SAS: Diagrama de Ramos-e-Folhas para tráfego maior que 7 milhões de minutos

O SAS apresenta também na mesma figura um diagrama de Caixas que utiliza a escala do diagrama de Ramos-e-Folhas. Neste diagrama, a Caixa apresenta uma simetria em relação à Média que está bem próxima da Mediana, conforme já observado na tabela 4.10

Concluindo a análise da amostra com valores maiores que 7 milhões de minutos verifica-se que esta amostra com 36 observações tem mais simetria, tem menor variabilidade e a hipótese de representar uma população com distribuição Normal é bem mais confiável do que a amostra anterior com 59 observações, ou seja, é recomendável trabalhar com a amostra de tráfego maior que 7 milhões de minutos.

Continuando o processo, analisa-se a seguir a amostra com os dias de tráfego menor 7 que milhões e maior que 3 milhões de minutos com base nos mesmos recursos do SAS anteriormente utilizados. Em decorrência da aplicação da função Distribution Analysis à amostra esta nova amostra obtém-se o resultado apresentado a seguir. Na tabela 4.15 é apresentada a relação de todas observações que caracterizam esta nova amostra com dias de tráfego menor que 7 milhões de minutos e maior que 3 milhões de minutos.

Tabela 4.15 SAS: Contagem de freqüência para tráfego menor que 7 milhões e maior que 3 milhões

Frequency Counts			
		Percents	
Value	Count	Cell	Cum
3262861.59	1	4.8	4.8
3268318.63	1	4.8	9.5
3804233.25	1	4.8	14.3
3856674.87	1	4.8	19.0
3898202.44	1	4.8	23.8
4066796.74	1	4.8	28.6
4092866.57	1	4.8	33.3
4143948.03	1	4.8	38.1
4267443.33	1	4.8	42.9
4278862.67	1	4.8	47.6
4733249.95	1	4.8	52.4
4812510.26	1	4.8	57.1
4846666.96	1	4.8	61.9
4864757.20	1	4.8	66.7
5089719.53	1	4.8	71.4
5275600.12	1	4.8	76.2
5587898.39	1	4.8	81.0
5625262.75	1	4.8	85.7
5711056.61	1	4.8	90.5
5782710.40	1	4.8	95.2
6454431.35	1	4.8	100.0

Essas novas observações formam uma nova amostra que será analisada e comparada com a amostra inicial apresentada na figura 4.2 e tabela 4.5

A tabela 4.16 apresenta os momentos de distribuição e outros parâmetros estatísticos.

Tabela 4.16 SAS: Momentos para tráfego menor que 7 milhões e maior que 3 milhões

Moments			
N	21	**Sum Weights**	21
Mean	4653527.22	**Sum Observations**	97724071.6
Std Deviation	869828.443	**Variance**	7.56602E11
Skewness	0.25392031	**Kurtosis**	-0.6490439
Coeff Variation	18.6918095	**Std Error Mean**	189812.128

Esta nova amostra é de 36 dias, com uma média de 4.635.527,22 minutos, o desvio padrão de 869.828,44 minutos e o coeficiente de variação de 18,69%. Comparando com a amostra inicial de 59 dias, verifica-se que o desvio padrão desta nova amostra é bem menor, quase 40% da amostra anterior de 2.122.557,77 minutos e que relação percentual de é 18,69% contra quase 31%. É possível afirmar, segundo Borror et al. (2006, p.166) que os dados desta nova amostra tem uma variabilidade bem menor que a da anterior, com uma grande concentração de valores em torno da média. Seguindo com a análise verifica-se que a estimativa de Assimetria apresentou o valor 0,25 caracterizam uma cauda assimétrica no sentido a direita da média. Entretanto a estimativa de Curtose foi de -0,65 que indica uma distribuição achatada.

A tabela 4.17 apresenta a Mediana com o valor de 4.733.250 minutos de tráfego menor do que o valor 7.855.738 da amostra anterior apresentado na tabela 4.2 devido a definição da amostra e bem próximo ao valor da Média de 4.653.527 com uma diferença de apenas 2,11%

Tabela 4.17 SAS: Parâmetros básicos para tráfego menor que 7 milhões e maior que 3 milhões

Basic Statistical Measures			
Location		**Variability**	
Mean	4653527	**Std Deviation**	869828
Median	4733250	**Variance**	7.56602E11

A tabela 4.18 apresenta as observações extremas relacionando as cinco observações com maiores valores que são menores do que as observações da amostra inicial apresentadas na tabela 4.3 e as cinco menores observações são maiores do que as da mesma, como o esperado pela própria característica na nova amostra escolhida

Tabela 4.18 SAS: Observações extremas para tráfego menor que 7 milhões e maior que 3 milhões

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
3262862	21	5587898	10
3268319	12	5625263	15
3804233	9	5711057	19
3856675	3	5782710	14
3898202	2	6454431	13

A tabela 4.19 apresenta os valores extremos desta nova amostra relacionando os cinco maiores valores que são menores do que os da amostra inicial apresentadas na tabela 4.4 e os cinco menores valores que são maiores do que as da mesma amostra, como o esperado pela própria característica da nova amostra escolhida

Tabela 4.19 SAS: Valores extremos para tráfego menor que 7 milhões e maior que 3 milhões

Extreme Values			
Lowest		Highest	
Order	Value	Order	Value
1	3262862	17	5587898
2	3268319	18	5625263
3	3804233	19	5711057
4	3856675	20	5782710
5	3898202	21	6454431

A tabela 4.20 apresenta os quantis da nova amostra, onde estão relacionados os percentis de maior relevância e os quartis.

Tabela 4.20 SAS: Quantis de maior relevância para tráfego menor que 7 milhões e maior que 3 milhões

Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	6454431
99%	6454431
95%	5782710
90%	5711057
75% Q3	5275600
50% Median	4733250
25% Q1	4066797
10%	3804233
5%	3268319
1%	3262862
0% Min	3262862

O 100º. percentil tem o valor 6.454.431 minutos de tráfego menor que o da amostra inicial apresentado na tabela 4.5. O 75º. percentil ou terceiro quartil tem o valor de 5.275.600 minutos, o 50º. percentil ou mediana tem o valor de 4.733.250 minutos de tráfego maior que o valor de 7.855.738 da amostra anterior e o 25º. percentil ou primeiro quartil tem o valor de 4.066.797 todos menores que os correspondentes quantis da amostra inicial apresentada na tabela 4.5. Entretanto o 0º. percentil ou valor mínimo da amostra de 3.262.862 minutos de tráfego maior que o valor de 1.053.702 da amostra inicial.

A figura 4.8 apresenta o Histograma do tráfego da amostra da tabela 4.15, onde os valores são automaticamente agrupados em intervalos de 600.000 de minutos de tráfego.

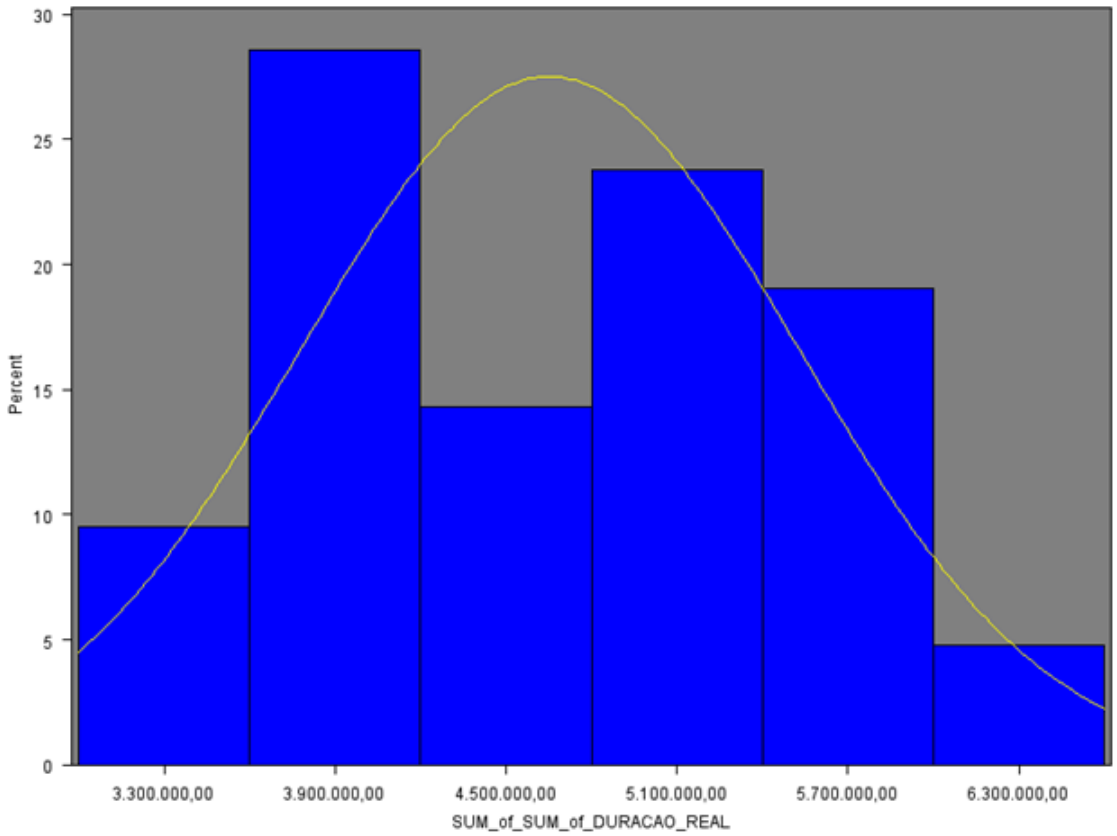


Figura 4.8 SAS: Histograma para tráfego menor que 7 milhões e maior que 3 milhões

O histograma desta amostra tem o formato mais próximo ao de um sino do que o da amostra inicial, apresentando uma simetria próxima a da distribuição Normal conforme indicado pelo parâmetro de Assimetria da tabela 4.16. Entretanto, existe uma diferença para o comportamento Normal que é existência de 2 picos relativos que confirma o resultado da Curtose da tabela 4.16. Além disto, o SAS apresenta também na mesma figura uma curva normal ajustada com base na média e o no desvio padrão da tabela 4.16

Um resultado muito relevante é do teste para Normalidade de Shapiro-Wilk apresentado na tabela 4.21 com um ρ–Value de 0,6231 bem mais próximo de 1 do que ρ–Value da amostra inicial que foi de 0,0001, ou seja, a hipótese de uma distribuição Normal é mais provável para a amostra de 21 dias

Tabela 4.21 SAS: Teste para normalidade para tráfego menor que 7 milhões e maior que 3 milhões

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.965055	**Pr < W**	0.6231
Kolmogorov-Smirnov	**D**	0.14286	**Pr > D**	>0.1500
Anderson-Darling	**A-Sq**	0.29805	**Pr > A-Sq**	>0.2500

A figura 4.9 apresenta o Diagrama Ramo-e-Folhas para o tráfego da amostra da tabela 4.15 numa escala de 10^6, onde a distribuição dos valores mostra uma concentração em torno do ramo $4x10^6$(11133), mas ainda com existência de dois picos nos ramos $4x10^6$(11133) e $5x10^6$(6678), conforme demonstrado pela Curtose da tabela 4.9 e mostrado no Histograma da figura 4.8.

```
Stem Leaf                  Freq.          Boxplot
   6 5                       1               |
   6                                         |
   5 6678                    4               |
   5 13                      2             +-----+
   4 7889                    4             *--+--*
   4 11133                   5             +-----+
   3 899                     3               |
   3 33                      2               |
   +----+----+----+----+----+----+
   Multiply Stem.Leaf by 10**+6
```

Figura 4.9 SAS: Diagrama de Ramos e-Folhas para tráfego menor que 7 milhões e maior que 3 milhões

Concluindo a análise onde a amostra inicial de 59 dias de tráfego foi dividida em três, sendo uma de 36 dias, outra com 21 dias e a última com dias que não foi analisada por não ter representatividade devido ao tamanho. A tabela 4.22 apresenta uma comparação da análise das três amostras.

Tabela 4.22 Comparação das amostras analisadas

PARÂMETRO	AMOSTRA		
	INICIAL	Tráfego > 7k	7k > Tráfego >3k
Tamanho (dias)	59	36	21
Média (Min)	6.783.375	8.320.740	4.653.527
Desvio Padrão (Min)	2.122.557	509.837	869.828
Coef. De Variação	31,29%	6,13%	18,69%
Assimetria	-0,801	-0,055	0,253
Curtose	-0,500	-0,516	-0,649
ρ– Value (S-W)	0,0001	0,6658	0,6231

O parâmetro do Coef. de Variação indica, segundo Borror et al. (2006, p. 166) que as amostras com 36 e 21 dias são mais homogêneas e com menor variabilidade. O parâmetro de Assimetria indica, segundo Schlotzhauer (2009, p.78) e Borror et al. (2006, p. 169) que as novas amostras tem um formato mais próximo do Normal do que amostra inicial. O parâmetro ρ–Value do teste de Shapiro-Wilk indica, segundo Schlotzhauer (2009, p.140) que a hipótese de amostra representar uma população com distribuição Normal é bem mais para as novas amostras com 36 e 21 dias do que para amostra inicial com 59 dias. Este resultado dá uma indicação de que a população analisada, no caso o tráfego gerado na cidade escolhida, tem comportamento diferenciado para grupo de dias com características assemelhadas. Analisando o gráfico da figura 4.4, percebe-se uma distribuição com 5 dias de alto tráfego seguido de dois dias tráfego mais baixo. Esta distribuição sugere análise do tráfego por dia de semana que será feito no tem seguinte.

4.2.3. Definindo as amostras para o processo de análise

Seguindo a sugestão do item anterior, obtêm-se no SAS a tabela 4.23 que relaciona os valores de tráfego da amostra de 59 dias com os dias da semana.

Tabela 4.23 SAS: Amostra de dia da semana para 59 dias

data	dia_da_semana	SUM_of_SUM_of_DURACAO_REAL
01/01/2010	Sexta	4.066.796,74
02/01/2010	Sábado	3.898.202,44
03/01/2010	Domingo	3.856.674,87
04/01/2010	Segunda	8.178.279,87
05/01/2010	Terça	8.518.410,89
06/01/2010	Quarta	8.296.551,71
07/01/2010	Quinta	7.958.836,52
08/01/2010	Sexta	8.006.706,62
09/01/2010	Sábado	5.089.719,53
10/01/2010	Domingo	4.864.757,20
11/01/2010	Segunda	8.277.379,69
12/01/2010	Terça	8.079.057,20
13/01/2010	Quarta	7.803.069,06
14/01/2010	Quinta	8.086.862,67
15/01/2010	Sexta	7.775.869,98
16/01/2010	Sábado	5.275.600,12
17/01/2010	Domingo	4.267.443,33
18/01/2010	Segunda	8.181.252,94
19/01/2010	Terça	7.855.737,56
20/01/2010	Quarta	7.549.287,16
21/01/2010	Quinta	7.924.079,60
22/01/2010	Sexta	7.299.360,03
23/01/2010	Sábado	4.812.510,26
24/01/2010	Domingo	3.804.233,25
25/01/2010	Segunda	8.201.779,52
26/01/2010	Terça	8.911.600,53
27/01/2010	Quarta	8.726.802,53
28/01/2010	Quinta	5.587.898,39
29/01/2010	Sexta	4.846.666,96
30/01/2010	Sábado	3.268.318,63
31/01/2010	Domingo	1.053.701,52
01/02/2010	Segunda	8.804.484,35
02/02/2010	Terça	8.893.631,09
03/02/2010	Quarta	8.938.565,57
04/02/2010	Quinta	8.862.706,93
05/02/2010	Sexta	8.846.371,34
06/02/2010	Sábado	6.454.431,35
07/02/2010	Domingo	5.782.710,40
08/02/2010	Segunda	9.264.814,64
09/02/2010	Terça	9.215.338,86
10/02/2010	Quarta	9.103.380,37
11/02/2010	Quinta	8.737.768,82
12/02/2010	Sexta	8.393.857,31
13/02/2010	Sábado	5.625.262,75
14/02/2010	Domingo	4.278.862,67
15/02/2010	Segunda	4.092.866,57
16/02/2010	Terça	4.143.948,03
17/02/2010	Quarta	7.279.958,07
18/02/2010	Quinta	8.393.862,23
19/02/2010	Sexta	8.141.334,43
20/02/2010	Sábado	5.711.056,61
21/02/2010	Domingo	4.733.249,95
22/02/2010	Segunda	8.414.801,49
23/02/2010	Terça	8.235.181,45
24/02/2010	Quarta	7.696.368,04
25/02/2010	Quinta	8.493.869,76
26/02/2010	Sexta	8.199.438,51
27/02/2010	Sábado	1.894.747,17
28/02/2010	Domingo	3.262.861,59

Os valores da tabela serão utilizados para definir a amostra do processo de análise do tráfego.

A partir de recursos do Excel obteve-se a figura 4.10 que apresenta a distribuição do tráfego por dia da semana com 41 dias úteis de tráfego, marcados com a cor sem graduação ou amarela e 18 dias de tráfego em fim de semana, marcados com uma graduação de cor mais forte na base ou vermelho.

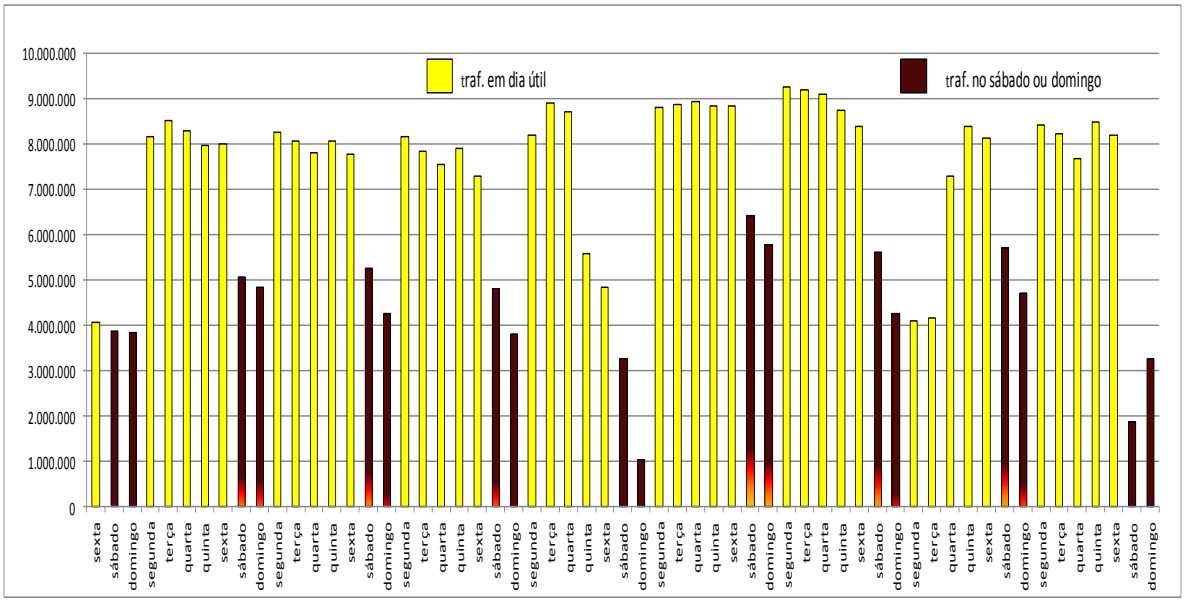


Figura 4.10 Distribuição do tráfego por dia de semana

Nesta figura observa-se que a distribuição de tráfego tem uma grande semelhança com as amostras de tráfego que analisamos inicialmente com base no valor do tráfego diário. A amostra de 36 dias obtida para o tráfego diário maior que 7 milhões de minutos está muito próxima da amostra de tráfego em dias úteis assinaladas em amarelo, com uma diferença de apenas 5 dias conforme apresentado tabela 4.24

Tabela 4.24 Amostra de dia da semana para Dia Útil

Dias	Data	Dia da semar	Tráfego
1	01/01/2010	Sexta	4.066.796,74
2	04/01/2010	Segunda	8.178.279,87
3	05/01/2010	Terça	8.518.410,89
4	06/01/2010	Quarta	8.296.551,71
5	07/01/2010	Quinta	7.958.836,52
6	08/01/2010	Sexta	8.006.706,62
7	11/01/2010	Segunda	8.277.379,69
8	12/01/2010	Terça	8.079.057,20
9	13/01/2010	Quarta	7.803.069,06
10	14/01/2010	Quinta	8.086.862,67
11	15/01/2010	Sexta	7.775.869,98
12	18/01/2010	Segunda	8.181.252,94
13	19/01/2010	Terça	7.855.737,56
14	20/01/2010	Quarta	7.549.287,16
15	21/01/2010	Quinta	7.924.079,60
16	22/01/2010	Sexta	7.299.360,03
17	25/01/2010	Segunda	8.201.779,52
18	26/01/2010	Terça	8.911.600,53
19	27/01/2010	Quarta	8.726.802,53
20	28/01/2010	Quinta	5.587.898,39
21	29/01/2010	Sexta	4.846.666,96
22	01/02/2010	Segunda	8.804.484,35
23	02/02/2010	Terça	8.893.631,09
24	03/02/2010	Quarta	8.938.565,57
25	04/02/2010	Quinta	8.862.706,93
26	05/02/2010	Sexta	8.846.371,34
27	08/02/2010	Segunda	9.264.814,64
28	09/02/2010	Terça	9.215.338,86
29	10/02/2010	Quarta	9.103.380,37
30	11/02/2010	Quinta	8.737.768,82
31	12/02/2010	Sexta	8.393.857,31
32	15/02/2010	Segunda	4.092.866,57
33	16/02/2010	Terça	4.143.948,03
34	17/02/2010	Quarta	7.279.958,07
35	18/02/2010	Quinta	8.393.862,23
36	19/02/2010	Sexta	8.141.334,43
37	22/02/2010	Segunda	8.414.801,49
38	23/02/2010	Terça	8.235.181,45
39	24/02/2010	Quarta	7.696.368,04
40	25/02/2010	Quinta	8.493.869,76
41	26/02/2010	Sexta	8.199.438,51

Nesta amostra os dias úteis com tráfego menor de 7 milhões estão assinalados em azul.

A amostra de 21 dias para tráfego menor que 7 milhões e maior que 3 milhões de minutos está bem próxima da amostra de tráfego para fim de semana assinaladas em azul, com uma diferença de apenas 3 dias conforme apresentado na tabela 4.25.

Tabela 4.25 Amostra de dia da semana para Sábado/Domingo

Dias	Data	Dia da semar	Tráfego
1	02/01/2010	Sábado	3.898.202,44
2	03/01/2010	Domingo	3.856.674,87
3	09/01/2010	Sábado	5.089.719,53
4	10/01/2010	Domingo	4.864.757,20
5	16/01/2010	Sábado	5.275.600,12
6	17/01/2010	Domingo	4.267.443,33
7	23/01/2010	Sábado	4.812.510,26
8	24/01/2010	Domingo	3.804.233,25
9	30/01/2010	Sábado	3.268.318,63
10	31/01/2010	Domingo	1.053.701,52
11	06/02/2010	Sábado	6.454.431,35
12	07/02/2010	Domingo	5.782.710,40
13	13/02/2010	Sábado	5.625.262,75
14	14/02/2010	Domingo	4.278.862,67
15	20/02/2010	Sábado	5.711.056,61
16	21/02/2010	Domingo	4.733.249,95
17	27/02/2010	Sábado	1.894.747,17
18	28/02/2010	Domingo	3.262.861,59

A diferença desta amostra para a amostra com tráfego menor que 7 milhões de minutos e maior que 3 milhões de minutos está nos 5 dias de a menos de tráfego de dias úteis menor que 7 milhões, marcados em azul na tabela 4.22 e dos 2 dias a mais de tráfego de fim de semana com tráfego menor que 3 milhões, marcados em azul na tabela 4.25

Esta semelhança indica que a análise da amostra de tráfego da cidade escolhida poderá ser realizada com mais precisão separando a amostra inicial em duas utilizando como parâmetro o tipo de dia: dia útil de segunda à sexta-feira e dia não útil para sábado e domingo. Esta hipótese pode ser confirmada através de uma análise mais detalhada para definição das novas amostras que chamaremos de amostra DU para dias úteis e amostra SD para os outros dias.

Iniciando pela a amostra DU percebe-se que os dias 1/Jan, 28/Jan, 29/Jan, 15/Fev e 16/Fev são pontos completamente fora da curva e não devem ser consideradas nas analises feitas da amostra DU que fica com 36 dias. A justificativa para o tráfego a menor de alguns desses dias é imediata. Dia 1/Jan é um dia feriado com comportamento completamente diferente de um dia útil. Os dias 15/Fev e 16/Fev são segunda-feira e terça-feira de Carnaval que não se comportam como dia útil. Os dias 28/Jan e 29/Jan serão analisados posteriormente para verificar se existe alguma anomalia. Analisando a amostra SD nota-se os dias 30/Jan, 31/Jan. 27/Fev e 28/Fev são pontos bem fora da curva. Neste caso a justificativa não é imediata e deve ser buscada posteriormente. A amostra SD fica com 14 dias. A figura 4.11 apresenta os dias que caracterizam as amostra DU e amostra SD como os seus respectivos pontos fora da curva.

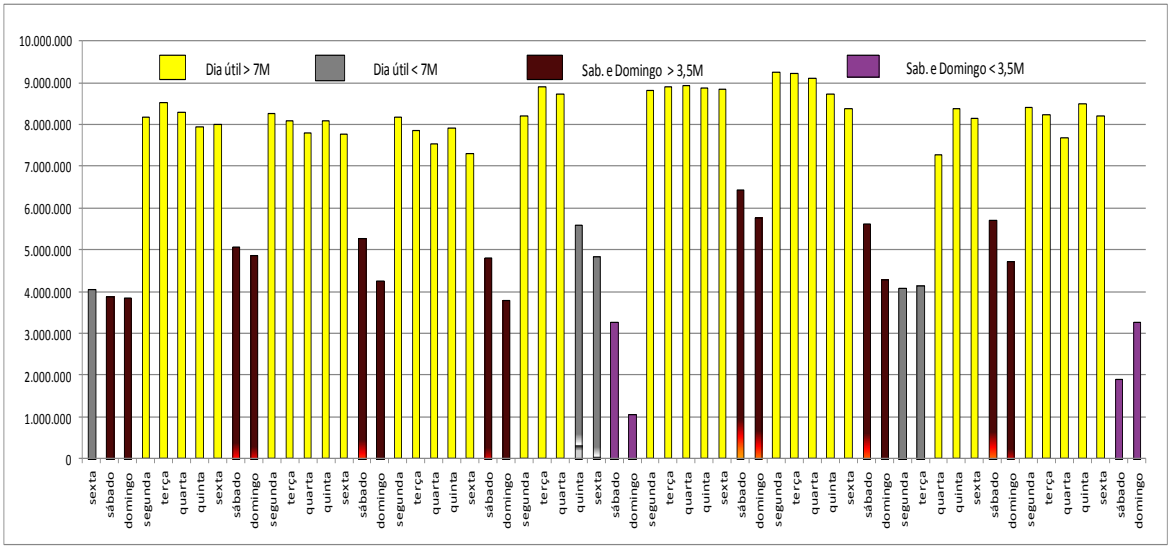


Figura 4.11 Distribuição do tráfego por dia da semana destacando os pontos fora da curvas

Em cor sem graduação ou amarelo estão marcados os dias da amostra DU e em graduação clara ou cinza estão marcados os dias úteis excluídos para uma melhor análise. Em graduação com a cor mais forte na base ou em vermelho estão marcados os dias da amostra SD e em graduação com a cor mais fraca na base ou em azul estão marcados os dias de Sábado e Domingos excluídos. Com a decisão de trabalhar com as amostras DU e SD no lugar de uma só, é importante fundamentar esta decisão submetendo cada umas destas amostras a um processo de análise utilizando a função Distribution Analysis do SAS nos moldes da análise de cada uma das amostras realizadas nos itens 4.2.1 e 4.2.2. No sentido de definir amostras mais homogêneas, um último ajuste foi feito com a retirada do dia

17/Fev que observou-se ser um ponto fora com a justificativa de ser quarta-feira de cinzas e com o comportamento diferente de um dia útil. Assim a amostra DU ficou com 35 dias ou eventos. O resultado obtido é pode ser observado na tabela 4.26 e nas figuras 4.12 e 4,13.

Tabela 4.26 Comparação de análise das amostras DU e SD

PARÂMETRO	AMOSTRA		
	INICIAL	DIAS	SÁBADO/DOMINGO
Tamanho	59	35	14
Média (Mn)	6.783.375	8.350.477	4.889.622
Desvio Padrão (Min)	2.122.557	484.571	816.667
Coef. De Variação	31,29%	5,80%	16,70%
Assimetria	-0,801	0,047	0,286
Curtose	-0,500	-0,572	-0,763
ρ– Value (S-W)	0,0001	0,7421	0,5840

O parâmetro do Coef. de Variação no valor de 5,80% para amostra DU e 16,70% para amostra SD, indica, segundo Borror et al. (2006, p. 166) que ambas amostras são mais homogêneas e com menor variabilidade que a amostra inicial. O parâmetro de Assimetria no valor de 0,047 para a amostra DU e 0,286 para a amostra SD indica, segundo Schlotzhauer (2009, p.78) e Borror et al. (2006, p. 169) e que as novas amostras tem um formato mais próximo do Normal do que amostra inicial. O parâmetro ρ–Value do teste de Shapiro-Wilk no valor de 0,7421 para amostra DU e 0,5840 para a amostra SD indica, segundo Schlotzhauer (2009, p.140) que a hipótese de as amostras representarem uma população com distribuição Normal é bem mais viável para as novas amostras DU e SD. A figura 4.12 mostra que a amostra DU tem uma distribuição próximo de uma Normal, mas com 2 picos relativos conforme indicado pelo valor da Curtose na tabela 4.24.

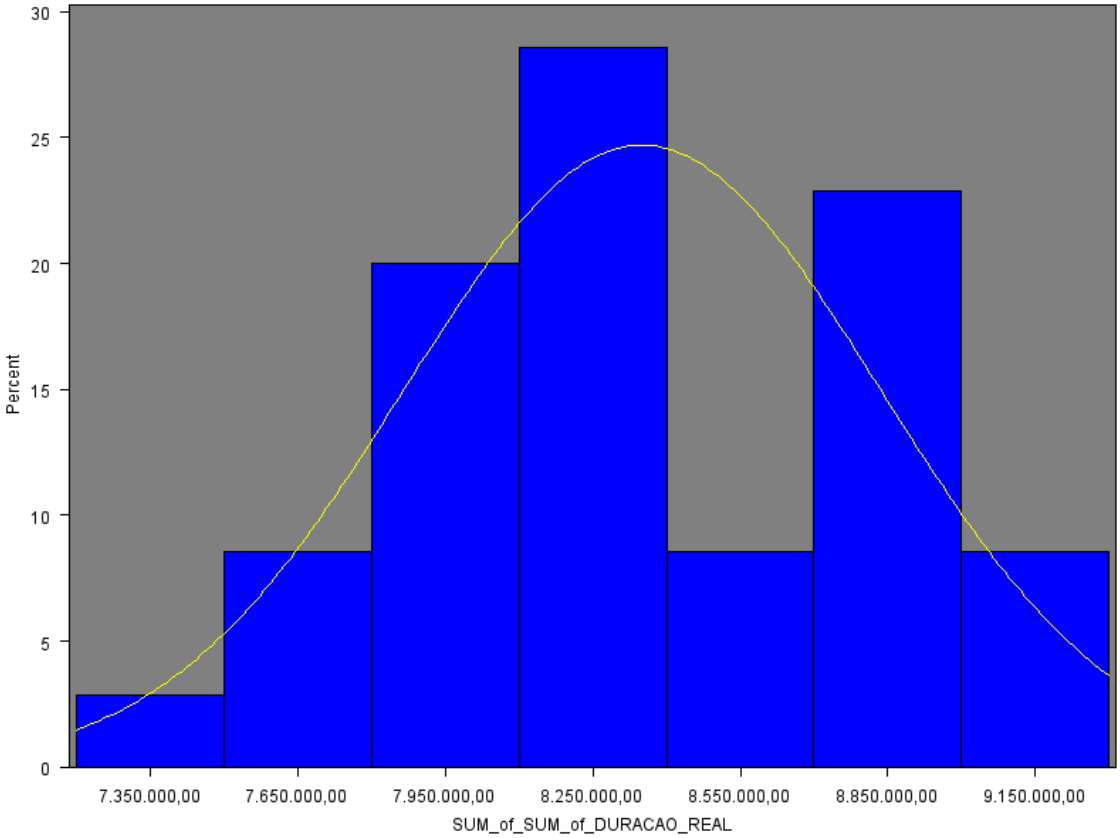


Figura 4.12 SAS: Histograma da amostra DU

A figura 4.13 mostra que a amostra SD tem uma distribuição próximo de achatada conforme indicado pelo alto valor de -0,763 da Curtose na tabela 4.24.

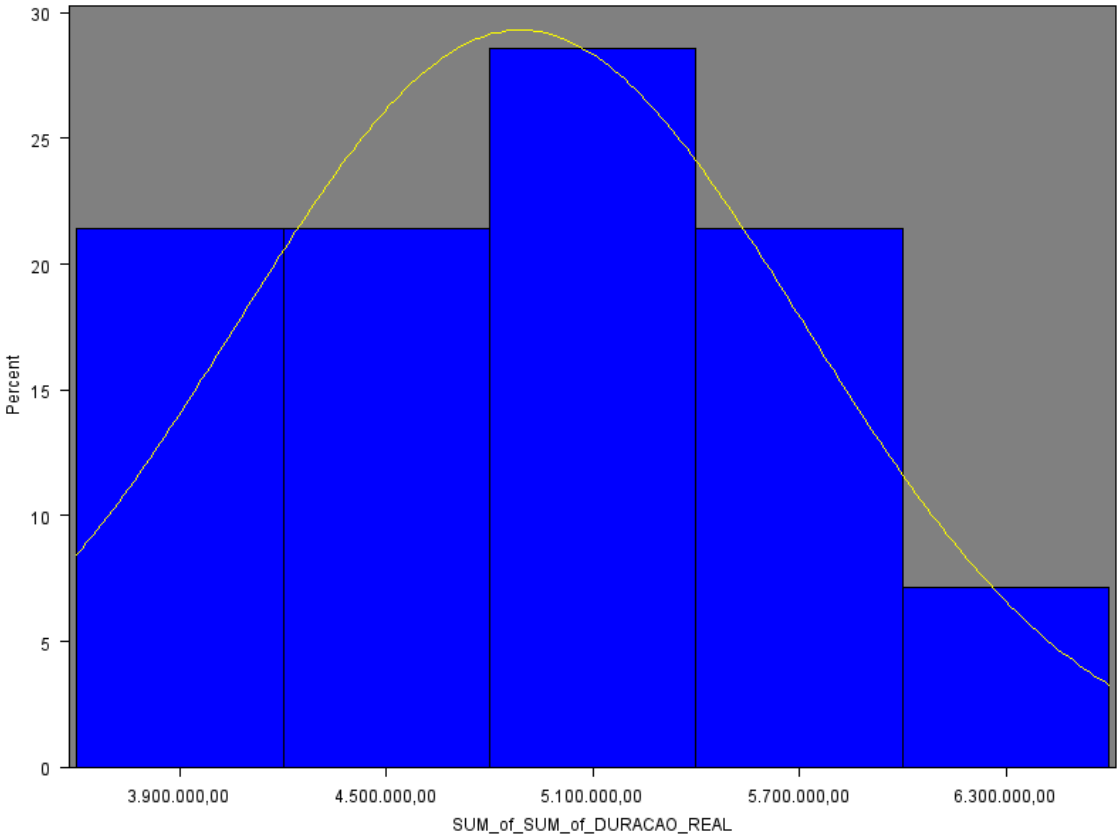


Figura 4.13 Histograma da amostra SD

Estes resultados sugerem a população analisada, no caso o tráfego gerado na cidade escolhida, tem comportamento diferenciado para grupo de dias com características assemelhadas e confirmam que as amostras DU e SD são uma melhor representação da população em estudo do que a amostra inicial com 59 dias.

5. ATUALIZAÇÃO DA BASE DADOS

Neste capítulo apresenta-se uma proposta de um processo para atualização da base de dados já definida no capítulo 3 com base nas amostras de tráfego definidas no capítulo 4. As amostras são tratadas com a utilização do SAS que garante a agilidade e confiabilidade do processo e do Excel para configuração de algumas tabelas e gráficos.

5.1. DEFININDO OS CRITÉRIOS

O processo de atualização da base de dados começa no inicio de cada mês pelo recebimento do tráfego diário coletado na cidade escolhida no caso desta dissertação, ou nas diversas filiais da empresa no caso prático. Neste processo é muito importante utilizar amostras sem anomalias que possam contaminar a base histórica de dados. Uma forma de garantir uma atualização segura é utilizar um intervalo de tolerância[26] mais abrangente que de acordo com Borror et al. (2006, p. 230) pode ser calculado pela técnica de Odeh e Owens (1980) ou utilizando a regra prática citada por Schlotzhauer (2009, p.137) que é uma aproximação do critério anterior. No caso desta dissertação foi utilizada a regra prática para dar mais agilidade ao processo. Dessa forma, o critério de atualização é o de

- +3

extremo inferior é o valor da média da amostra

padrão da amostra(S) e o limite superior é o valor da média da amostra

o valor do desvio padrão da amostra(S). Caso o valor do tráfego esteja fora do intervalo não será considerado na atualização e será alvo de uma análise criteriosa que será comentada e exemplificada no final deste capítulo. Esse garante que, no caso de uma distribuição normal, 99% dos valores estão dentro do intervalo.

O critério escolhido deverá ser aplicado em separado para cada tipo de amostra. Se o tráfego for de um dia útil, a comparação se dará com a amostra DU, mas se for de um sábado ou domingo a comparação será com a amostra SD. Se o tráfego for de um dia feriado será tratado de uma forma diferenciada que será proposta no capitulo 6 desta dissertação

[26] Intervalos de tolerância são intervalos nos quais esperamos que estejam contidos uma porcentagem dos valores populacionais

5.2. ATUALIZANDO A BASE DE DADOS

A demonstração de aplicação do critério de seleção escolhido é feita com a utilização parcial da amostra DU. A idéia é utilizar os dias de Janeiro da amostra DU como base de dados e aplicar o critério de seleção para os primeiros dias de Fevereiro da mesma amostra, simulando o recebimento de novos dias de tráfego em processo prático. Sendo assim na tabela 5.1 é apresentada a base de dados a ser utilizada como demonstração do processo de atualização.

Tabela 5.1 Amostra DU para Janeiro de 2010

Dias	Data	Dia da seman	Tráfego
1	04/01/2010	Segunda	8.178.279,87
2	05/01/2010	Terça	8.518.410,89
3	06/01/2010	Quarta	8.296.551,71
4	07/01/2010	Quinta	7.958.836,52
5	08/01/2010	Sexta	8.006.706,62
6	11/01/2010	Segunda	8.277.379,69
7	12/01/2010	Terça	8.079.057,20
8	13/01/2010	Quarta	7.803.069,06
9	14/01/2010	Quinta	8.086.862,67
10	15/01/2010	Sexta	7.775.869,98
11	18/01/2010	Segunda	8.181.252,94
12	19/01/2010	Terça	7.855.737,56
13	20/01/2010	Quarta	7.549.287,16
14	21/01/2010	Quinta	7.924.079,60
15	22/01/2010	Sexta	7.299.360,03
16	25/01/2010	Segunda	8.201.779,52
17	26/01/2010	Terça	8.911.600,53
18	27/01/2010	Quarta	8.726.802,53

Aplicando a função Distribution Analysis do SAS à amostra DU de Janeiro obtém-se o resultado para os parâmetros estatísticos da tabela 5.2.

Tabela 5.2 SAS: Momentos da amostra DU para Janeiro 2010

Moments			
N	18	**Sum Weights**	18
Mean	8090606.89	**Sum Observations**	145630924
Std Deviation	389240.035	**Variance**	1.51508E11
Skewness	0.1995252	**Kurtosis**	0.59999134
Coeff Variation	4.81101158	**Std Error Mean**	91744.7561

Utilizando a Média e o Desvio padrão da tabela 5.2 calcula-se o intervalo de tolerância do critério de seleção para os aos futuros valores de tráfego. A tabela 5.3 apresenta o intervalo de tolerância.

Tabela 5.3 Intervalo de tolerância

Intervalo de tolerância da amostra DU		
Limite Inferior	Média	Limite Superior
6.922.346,77	8.090.606,89	9.258.867,01

Apesar do processo de coleta definido no item 3.1 gerar novos dados de tráfego diariamente, a proposta dessa dissertação é de realizar a comparação do tráfego com a amostra a cada sete de dias de coleta para garantir uma amostra mais representativa do mês e evitar um retrabalho diário. A idéia de sete dias é obter sempre 5 dias de tráfego útil mais sábado e domingo. Embora o processo de análise seja semanal, é necessário que a cada coleta diária, o tráfego seja processado no SAS para gerar novos dados tráfego por chamada tal qual apresentado na tabela 5.4.

Tabela 5.4 Tráfego por chamada

DIA_MES	CENTRAL	HORA (intervalo)	DURACAO_REAL (Min)
28/fev	CD1	10	0.65
28/fev	CD1	10	0.13
28/fev	CD1	9	4.6
28/fev	CD1	9	9.3
28/fev	CD1	10	3.42
28/fev	CD1	10	1.05
28/fev	CD1	10	7.82
28/fev	CD1	10	1.55
28/fev	CD1	10	1.17
28/fev	CD1	10	0.6

Esta tabela contém para cada chamada, os dados de central, intervalo de hora e duração da chamada. Outro dado importante é resumo do tráfego por intervalo de hora apresentado na tabela 5.5.

Tabela 5.5 Tráfego por intervalo de hora

Data	Dia da semana	Hora (intervalo)	Duração Real (min)
28/02/2010	Domingo	0	91944,18
28/02/2010	Domingo	1	123524,54
28/02/2010	Domingo	2	59667,73
28/02/2010	Domingo	3	58788,36
28/02/2010	Domingo	4	45426,26

O processo de atualização da base de dados tem uma etapa ao final da primeira semana do mês de quando a equipe responsável pela análise de tráfego receber uma amostra com o tráfego dos sete primeiros. Nesta demonstração do processo utiliza-se uma amostra DU com o tráfego de dos 5 primeiro dias úteis de Fevereiro de 2010 que são apresentados na tabela 5.6

. Tabela 5.6 Tráfego dos primeiros dias úteis de Fevereiro

Data	Dia da Semana	Duração (Min)
01/02/2010	Segunda	8.804.484,35
02/02/2010	Terça	8.893.631,09
03/02/2010	Qurta	8.938.565,57
04/02/2010	Quinta	8.862.706,93
05/02/2010	Sexta	8.846.371,34

Analisando cada valor de tráfego da tabela 5.6 e comparando com o critério da tabela 5.3, conclui-se que todos os valores que estão dentro do intervalo de tolerância. Dessa forma a base de dados DU é atualizada com mais 5 dias atingindo o tamanho de 23 dias Esta amostra será novamente atualizada a cada 7 dias de tráfego.

Aplicando a função Distribution Analysis do SAS à nova amostra DU de Janeiro/Fevereiro obtém-se como resultado os parâmetros estatísticos da tabela 5.7.

Tabela 5.7 SAS:Parâmetros estatísticos da amostra DU de Janeiro/Fevereiro

Moments			
N	23	**Sum Weights**	23
Mean	8259855.8	**Sum Observations**	189976683
Std Deviation	474706.342	**Variance**	2.25346E11
Skewness	-0.0314654	**Kurtosis**	-0.8659602
Coeff Variation	5.74715047	**Std Error Mean**	98983.1147

Um bom tamanho da amostra de um processo rodando em operadoras de Telecom é o de 365 dias, o período de ano, ou seja, a amostra DU deve ter 260 dias descontando os

feriados e a amostra SD com 105 dias também com descontos dos feriados. Quando a amostra completar o período de 12 meses, por exemplo, de Janeiro a Dezembro de 2010 e com a continuação do processo de coleta, os novos dias de tráfego de Janeiro de 2011 substituirão os dias de tráfego de Janeiro de 2010 serão descartados. A proposta é manter sempre uma base de dados atualizada com os últimos 12 meses, evitando que um histórico muito grande possa mascarar as novas tendências nascentes.

Os dias feriados têm valores de tráfego bastante distinto entre eles, bem diferente dos dias úteis e mais próximos dos valores tráfego de um Sábado e Domingo. Na amostra inicial de 59 dias apresentada na figura 4.2 e tabela 4.5 existem 4 dias feriados: 1/Janeiro dia da Confraternização Universal; 15,16 e 17/Fevereiro dias de carnaval. Os 3 primeiros apresentam valores de tráfego próximos dos valores de um Domingo. O quarto dia, a Quarta-feira de Cinzas apresentou um valor próximo ao de um dia útil. De um modo geral o tráfego em dias feriados é bem menor que um dia normal, embora haja exceções com o Dia das Mães cujo o volume de tráfego é maior que o tráfego de muitos dias úteis, mesmo sendo um Domingo. Considerando esta diversidade de comportamento a proposta deste processo é considerar os dias de feriados relacionados com dias das amostras DU ou SD e armazenados em um arquivo de fácil acesso. Exemplificando, os dias 1/Janeiro, Segunda e Terça-feira de Carnaval e Quarta-feira de cinzas são classificados como um dia de amostra DU e o tráfego relacionado com percentual de tráfego de um dia útil. Um feriado que caia em Sábado ou Domingo deve ser classificado como tráfego da amostra SD e relacionado com um percentual de tráfego de um Sábado ou Domingo normal. Este arquivo com a classificação dos dias feriados será utilizado no capítulo 6 na estimativa de tráfego.

5.3. ANÁLISE DE ANOMALIA

Voltando ao assunto do dia cujo valor de tráfego está fora do intervalo de comparação, o modo de lidar com este problema é fazer uma análise top down para descobrir se existe alguma anomalia. Considerando o tráfego mensal de uma operadora de Telecom, o primeiro passo é fazer a análise do tráfego desse dia com abertura por filial. Uma vez identificada a filial, o processo continua para mapear em quais cidades foi percebido um comportamento diferenciado do tráfego. Caso este fato seja restrito a uma cidade, a análise pode descer ao nível de central. Caso o comportamento diferenciado do

tráfego seja observado em uma ou mais filial, ou em um grupo de cidades a análise pode ser feita por intervalo de hora. Exemplificando o processo proposto, utiliza-se a base de dados da amostra inicial de 59 dias e dentro dela o dia 27/Fevereiro, um Sábado que apresenta um dos menores valores de tráfego. O primeiro passo da análise top down é descobrir em quais das 42 centrais da localidade ocorre o fato do tráfego do dia 27/Fevereiro ser bem menor do que nos outros dias. Dentro da amostra SD, seleciona-se uma sub-amostra somente com os dias de Fevereiro conforme apresentada na tabela 5.8.

Tabela 5.8 Amostra SD para Fevereiro

Dias	Data	Dia da semana	Tráfego
1	06/02/2010	Sábado	6.454.431,35
2	07/02/2010	Domingo	5.782.710,40
3	13/02/2010	Sábado	5.625.262,75
4	14/02/2010	Domingo	4.278.862,67
5	20/02/2010	Sábado	5.711.056,61
6	21/02/2010	Domingo	4.733.249,95
7	27/02/2010	Sábado	1.894.747,17

A idéia de utilizar uma amostra somente com dias de Fevereiro é para tornar a demonstração da análise menos complexa e mais rápida.

Aplicando a função Distribution Analysis do SAS obtêm-se a distribuição diária do tráfego para cada central conforme apresentado na figura 5.1 e tabela 5.9 para a central CD1.

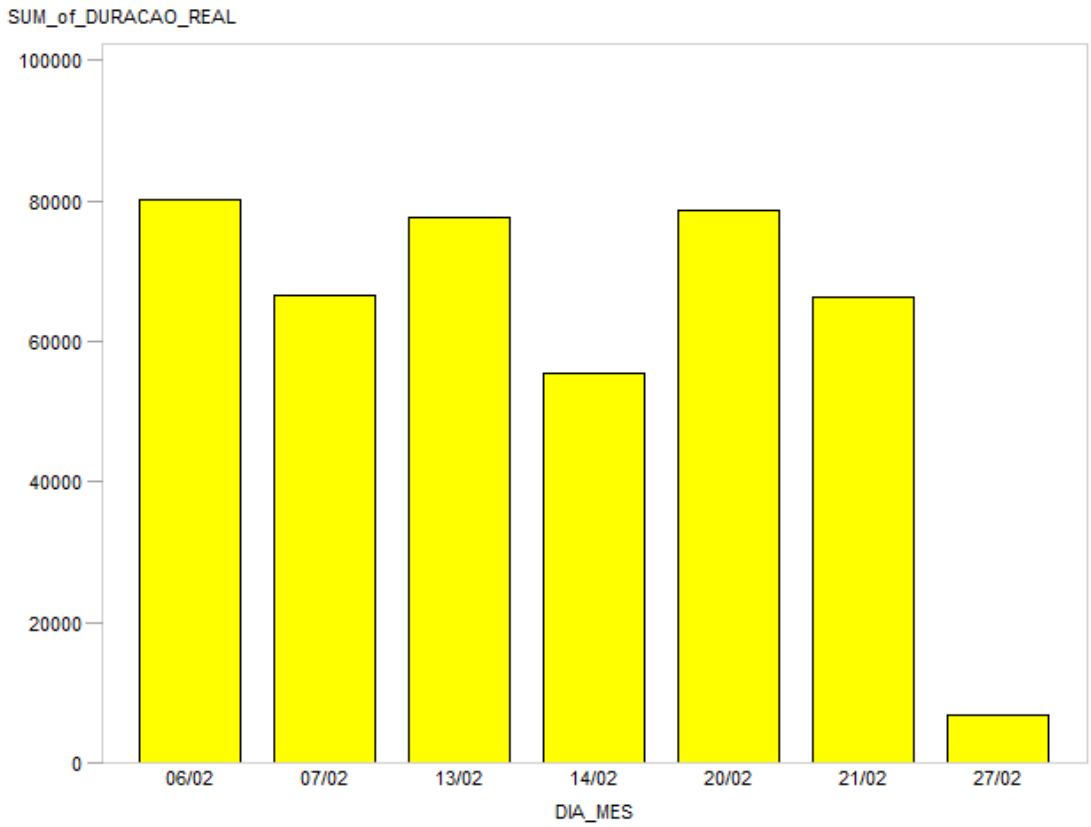


Figura 5.1 SAS: Tráfego da central CD1 em diversos dias da amostra SD

Tabela 5.9 Tráfego da amostra SD para central CD1

DIA_MES	SUM_of_DURACAO_REAL
06/02	80070.8
07/02	66510.7
13/02	77547
14/02	55570.4
20/02	78580.2
21/02	66275.7
27/02	6881.73

Observa-se que o tráfego da central CD1 para o dia 27 é da ordem de 12% do valor do menor tráfego dos outros dias. A análise do resultado do SAS para as 42 centrais da amostra inicial indica que este comportamento do tráfego do dia 27/Fevereiro é observado em 34 das centrais, ou seja, em 80% das centrais. Dada a abrangência do efeito, conclui-se, numa avaliação inicial que o comportamento diferenciado do tráfego não é causado por nenhuma central especifica. É necessário aprofundar análise dentro das centrais e como exemplo, é utilizado a central CD1, já apresentada na figura 5.1e com a utilização do SAS,

obtêm-se uma da distribuição do tráfego do dia 27/Fevereiro por hora que é mostrado na figura 5.2.

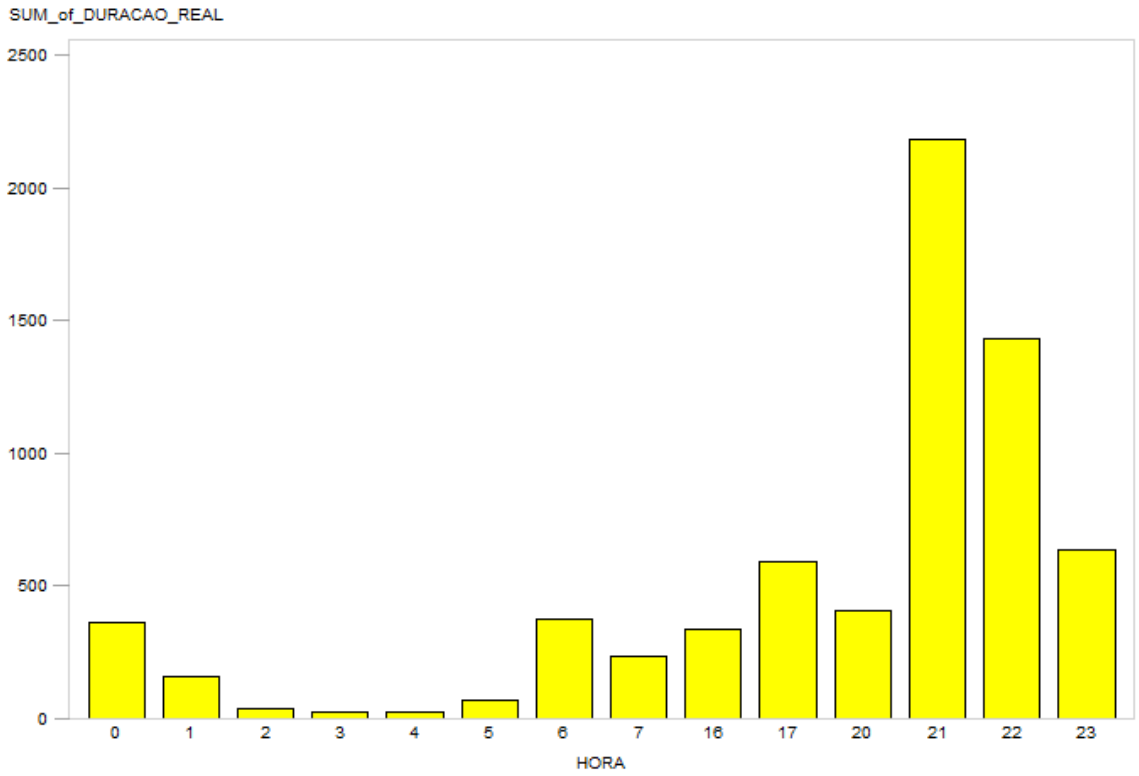


Figura 5.2 SAS: Distribuição do tráfego da central CD1 por intervalo de hora

Nesta figura observa-se que no 27/Fevereiro não existem valores de tráfego para os intervalos de hora: 8H até 15H; 18H e 19H e o mesmo efeito é observado em mais 31 centrais conforme apresentado na figura 5.3.

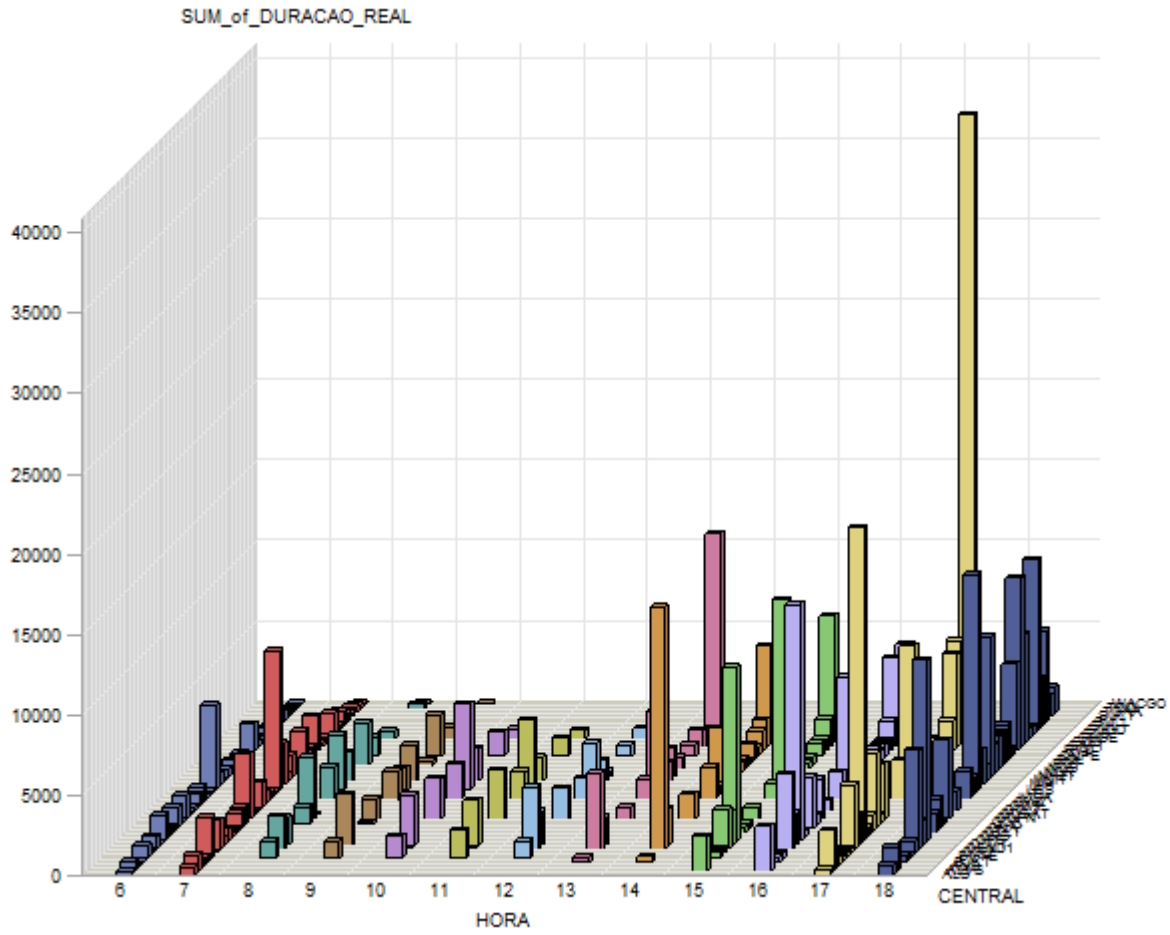


Figura 5.3 SAS: Distribuição horária do tráfego para todas as centrais no dia 27/Fevereiro

Esta figura indica que no intervalo de 8H às 15H, não existiu tráfego medido para um grande número de centrais. Entretanto, este mesmo comportamento não é observado em outros dias, como o dia 6/Fevereiro mostrado na figura 5.4.

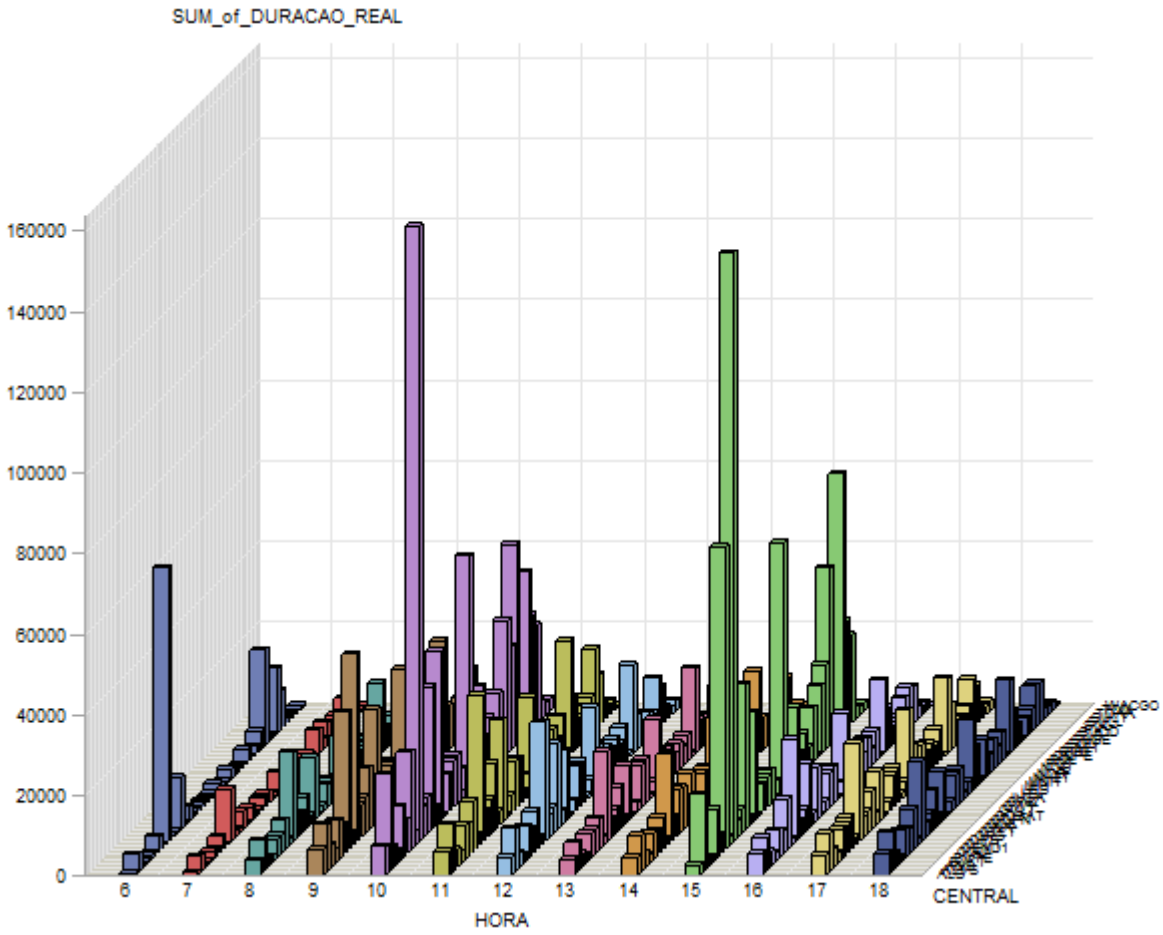


Figura 5.4 SAS: Distribuição horária do tráfego para todas as centrais no dia 6/Fevereiro

Esta figura indica que no dia 6/Fevereiro existe tráfego distribuído por todos os intervalos de hora sem anomalia observada no dia 27/Fevereiro.

Na figura 5.5 tem-se uma visão geral da distribuição do tráfego por intervalo de horas em todos os dias da amostra SD de Fevereiro.

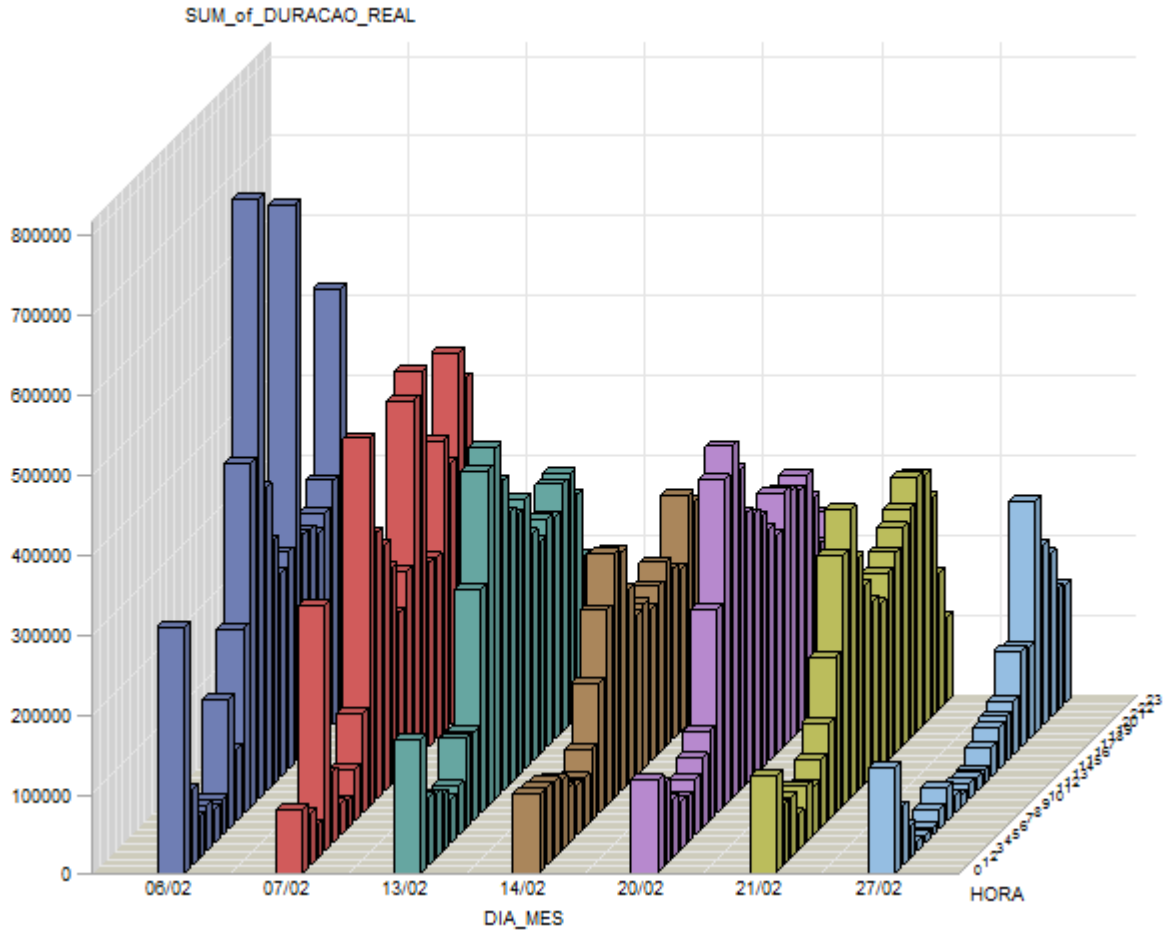


Figura 5.5 SAS: Distribuição horária do tráfego para a amostra SD de Fevereiro

Nesta figura observa-se que a distribuição do tráfego entre os intervalos de 5H até 19H tem o formato bem mais achatado que os de todos outros dias. Como resultado desta análise é possível afirmar existe uma anomalia no dia 27/Fevereiro que provocou uma perda no tráfego da ordem de 3,0 milhões de minutos[27], estimado com base na diferença para a Média da amostra SD.

Concluindo a análise, gera-se um relatório que deverá ser enviado à Gerencia de Rede da filial onde está localizada a cidade em estudo para que a etapa do processo de coleta de tráfego onde ocorreu a anomalia seja identificada o mais rápido possível. Perceba que o processo proposto pode detectar anomalias no máximo uma semana após a ocorrência da mesma, sendo bem mais ágil do que maioria dos processos em execução nas empresa que levam de 40 a 60 dias para detectar anomalias conforme descrito na introdução dessa dissertação.

[27] Na tabela 5.8 o tráfego para o dia 27/Fevereiro é de 1.894.747,17 minutos de tráfego e a Média da amostra SD na tabela 4.24 é 4.889.622 minutos de tráfego. Esta diferença é da ordem de 3,0 minutos de tráfego

5.4. PROCESSO DE ATUALIZAÇÃO DA BASE DE DADOS

Concluindo, afirma-se que o processo descrito neste capítulo e resumido na figura 5.6 é uma boa ferramenta para gerenciamento do tráfego telefônico de voz em processo de tarifação, mas não tem a pretensão de ser o único. É importante entender que a descrição aqui proposta é de um processo geral que ao ser implantado deve ser customizado, ou seja, adaptado as características de cada empresa, bem como às suas prioridades

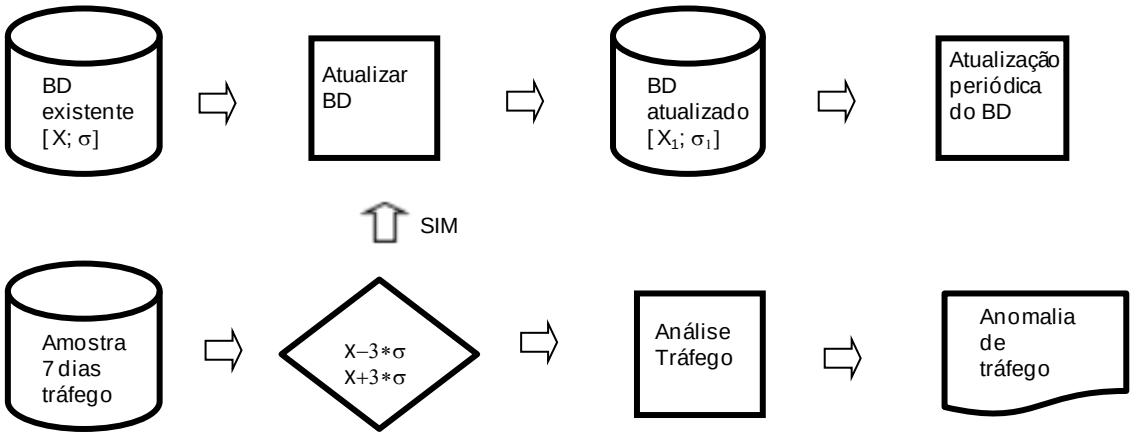


Figura 5.6 Processo para atualização do banco de dados

6. ESTIMATIVA DE TRÁFEGO

Neste capítulo é apresentada uma proposta de um processo para a estimativa do tráfego mensal com base em amostra dos primeiros dias de tráfego do mês corrente. As amostras são tratadas com a utilização do SAS que garante a agilidade e confiabilidade do processo e do Excel para configuração de algumas tabelas e gráficos.

6.1. VALIDANDO A AMOSTRA

Considerando que o processo de atualização da base dados está sendo executado conforme descrito no capítulo anterior, a etapa seguinte para completar o processo de gerenciamento do tráfego é a estimativa mensal do tráfego. Esta previsão pode antecipar em até 30 dias o conhecimento da receita mensal de voz de longa distância do período, além de detectar com antecedência anormalidades no processo de coleta das informações de tráfego telefônico.

O processo de cálculo da estimativa inicia-se com o recebimento de uma amostra de tráfego dos 7 primeiros dias do mês corrente. Os dias desta amostra são testados pelo *S- +3 conforme descrito no capítulo anterior e aqueles que estiverem dentro do intervalo são a base para da previsão. Em seguida é feita a identificação da quantidade de dias no mês e a classificação do dia. Cada dia pode ser classificado em Dia Útil, Sábado e Domingo ou Feriado. Nesta proposta optou-se por considerar o comportamento do tráfego de um Sábado semelhante ao de Domingo porque a análise da amostra inicial de 59 dias realizada no capítulo anterior apontou para esta característica. Entretanto, com uma amostra maior pode ser que a analise aponte para outra característica e a amostra SD seja na verdade separada em duas: uma para dias de Sábado e outra para dias de Domingo. Cada dia Feriado deve ter um tratamento diferenciado. Se for um Feriado nacional o impacto deve ser considerado em todas filiais assim como Feriados estaduais e municipais devem impactar somente as respectivas filiais e municípios. Um Feriado pode cair durante a semana ou no fim-de-semana e como tal o comportamento de tráfego pode ser equivalente à média da amostra DU ou amostra SD. Se o Feriado ocorrer durante a semana de trabalho, o projetista deverá classificar o comportamento de tráfego

em função da amostra DU estipulando um valor para relação. Por exemplo, no mês de Fevereiro da amostra inicial, a Segunda e Terça-feira de Carnaval podem ser relacionadas com 50% do tráfego de um dia útil, isto significa que o tráfego previsto para esse dia será 50% do valor da média da amostra DU. Na última etapa do processo de estimativa, efetua-se o cálculo do tráfego do mês corrente com base nos dados recebidos da quantidade e da classificação dos dias e do tráfego correspondente aos 7 primeiro dias do mês. Este processo deve ser repetido com os próximos 7 dias do mês, ou seja com os dias 8,9,10,11,12,13 e 14 para confirmar ou atualizar a estimativa. Este processo fica mais bem explicado com um exemplo onde o tráfego previsto para Fevereiro de 2010 será calculado com base nas amostras DU e SD obtidas com os primeiros 7 dias de tráfego do mês de Fevereiro de 2010 apresentados na tabela 6.1.

. Tabela 6.1 Tráfego dos primeiros dias de fevereiro

Data	Dia da semana	Duração (min)
01/02/2010	Segunda	8.804.484,35
02/02/2010	Terça	8.893.631,09
03/02/2010	Quarta	8.938.565,57
04/02/2010	Quinta	8.862.706,93
05/02/2010	Sexta	8.846.371,34
06/02/2010	Sábado	6.454.431,35
07/02/2010	Domingo	5.782.710,40

A próxima etapa da estimativa é testar se os valores de tráfego estão dentro do intervalo das amostras DU e SD e aqui se considera que o teste para o dia Útil já foi efetuado com êxito no capítulo anterior. Para o teste é necessário calcular a Média e Desvio Padrão para amostra SD somente com os dias de Janeiro obtida na com a aplicação da função Distribution Analysis do SAS e apresentada na tabela 6.2.

Tabela 6.2 SAS: Parâmetros estatísticos para a amostra SD para Janeiro

Moments			
N	8	**Sum Weights**	8
Mean	4483642.63	**Sum Observations**	35869141
Std Deviation	596668.865	**Variance**	3.56014E11
Skewness	0.03536592	**Kurtosis**	-2.0641404
Coeff Variation	13.3076812	**Std Error Mean**	210954.3

Utilizando a Média e o Desvio padrão da tabela 6.2 calcula-se o intervalo de tolerância do critério de seleção para os aos futuros valores de tráfego. A tabela 6.3 apresenta o intervalo de tolerância.

Tabela 6.3 Intervalo de tolerância para a amostra SD de Janeiro

Intervalo de tolerância da amostra SD Janeiro		
Limite Inferior	Média	Limite Superior
2.693.636,04	4.483.642,63	6.273.649,23

A análise envolve dois novos valores de tráfego, 6.454.431 minutos para o dia 6/Fev e 5.782.710 minutos para o dia 7/Fev e o intervalo da amostra SD com limite inferior de 2.693.636,04 minutos e superior de 6.273.649,23 minutos. O valor de 6/Fev está fora do intervalo e portanto, a estimativa será feita com o valor de 5.782.710 minutos para dias de Sábado e Domingo.

6.2. ESTIMANDO O TRÁFEGO

A estimativa do tráfego para o mês de Fevereiro é calculada com base na amostra DU e SD dos primeiros 7 dias deste mês apresentada na tabela 6.1 para os valores que estiverem dentro dos intervalos de tolerância já calculados e apresentados nas tabelas 5.3 e 6.3. O parâmetro utilizado para estimar o tráfego é a média amostral que no caso da amostra SD é 5.782.710 porque foi o único que passou no teste do intervalo de tolerância, mas no caso da amostra DU é a Média calculada com a aplicação da função Distribution Analysis do SAS e apresentada na tabela 6.4.

Tabela 6.4 SAS: Parâmetros estatísticos da amostra DU de Fevereiro

Moments			
N	5	**Sum Weights**	5
Mean	8869151.86	**Sum Observations**	44345759.3
Std Deviation	50398.095	**Variance**	2539967975
Skewness	0.21678228	**Kurtosis**	0.0086717
Coeff Variation	0.5682403	**Std Error Mean**	22538.7133

O cálculo do tráfego e a classificação dos dias estão representados na tabela 6.5 onde

.

Tabela 6.5 Classificação dos dias e estimativa de tráfego

Dia	Qtde	Média	% Dial	Tráfego
Útil	17	8.869.151,86		150.775.581,62
Sáb/Dom	6	5.782.710,00		34.696.260,00
13/Sábado Carnaval	1	2.891.355,00	50%	2.891.355,00
14/Domingo Carnaval	1	2.891.355,00	50%	2.891.355,00
15/Segunda Carnaval	1	4.434.575,93	50%	4.434.575,93
16/Terça Carnaval	1	4.434.575,93	50%	4.434.575,93
17/Quarta Cinzas	1	4.434.575,93	50%	4.434.575,93
Total	28			204.558.279,41

se observa que o mês de Fevereiro de 2010 foi classificado com tendo 17 dias úteis, 6 dias de Sábado e Domingo e 5 dias feriados. Estima-se que cada dia útil tenha o tráfego de 8.869.151,86 minutos que representa a média dos primeiros dias do mês corrente e cada Sábado e Domingo tenha o tráfego de 5.782.710 minutos que representa a média dos primeiro dias com expurgo do dia que não passou no teste - +3
Feriado estimou-se que o tráfego seria equivalente ao tráfego do dia sem o feriado com uma redução de 50%. Assim o dia 15/Fev, uma Segunda-feira, se não fosse o Carnaval, teria o tráfego de 8.869.151,86 minutos, mas como existe o efeito Carnaval o tráfego a ser considerado é de 4.434.575,93 minutos.

Este critério pode ser utilizado quando não se dispõe do histórico do ano anterior. No caso de um processo rodando em uma empresa e com uma amostra anual e um histórico maior que 1 ano, a classificação destes dias deverá ser feita com base no histórico de tráfego de cada filial. Ainda, na tabela 6.5, o valor total do tráfego estimado para Fevereiro foi de 204.558.279,41 minutos. Comparando a estimativa com o valor do tráfego realmente coletado de Fevereiro, no valor de 199.895.730,35 minutos, observa-se uma diferença de 4.662.549,06 da ordem de 2,3% do valor real. Esta diferença pode representar dois efeitos, sendo o primeiro um reflexo da precisão do cálculo que é reflexo imediato do tamanho e da qualidade da amostra e do histórico utilizado. Com amostras de 365 dias e com a experiência do projetista, esta diferença deve cair para valores bem mais baixos. O outro efeito importante é o impacto das anomalias que ocorrem nos diversos elementos de rede responsáveis pela coleta e transmissão dos dados dos CDRs e pela operação da rede

telefônica. Neste exemplo, parte da diferença de 4.662.549,06 pode justificada pela anomalia ocorrida dia 27/Fev com uma perda no tráfego da ordem de 3,0 milhões de minutos.

6.3. PROCESSO DE ESTIMATIVA DO TRÁFEGO MENSAL

Concluindo este capítulo, mostra-se a figura 6.1 um resumo do processo de estimativa mensal do tráfego onde é importante manter o foco do processo. Se o contexto da estimativa é uma empresa regional, o cálculo deverá ser feito com tráfego dos municípios/ cidades ou regiões de abrangência. Se o contexto for uma empresa nacional ou internacional o processo de estimativa deverá focar nas filais, grupo de filiais ou países o cálculo deverá ser da área.

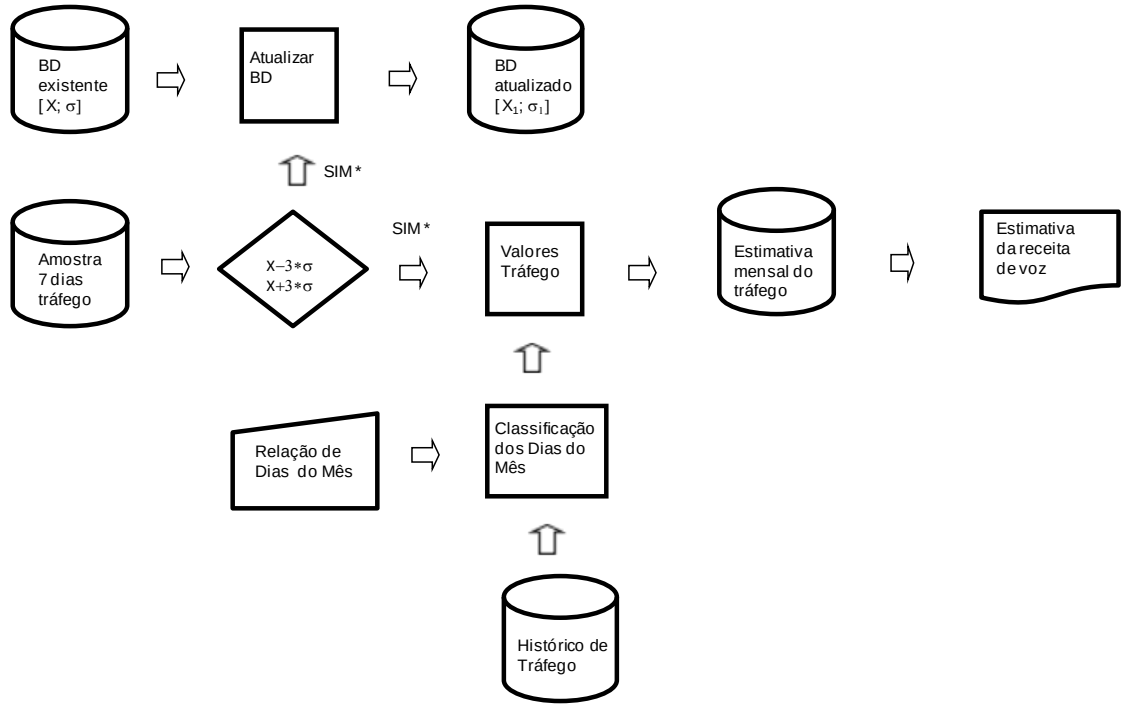


Figura 6.1 Processo para estimativa do tráfego mensal

É importante sempre lembrar que o principal objetivo da estimativa de tráfego é antecipar os resultados mensais que permitam diagnósticos de anomalias e novas tendências que demandam ações gerenciais corretivas e preventivas

7. CONCLUSÕES

Neste trabalho foi apresentado um processo para gerenciamento do tráfego telefônico de voz responsável pela geração da maior fonte de receita das Operadoras de Telecom Fixa sendo também uma receita bastante relevante nas Operadoras Móveis. Esta receita, hoje no Brasil de acordo com dados de Teletime(2009), é da ordem de mais R$ 50 bilhões no ano e demanda soluções robustas, agéis e muito confiáveis porque qualquer pequeno desvio produz impactos de grandes proporções: um desvio de 1% causa um impacto de R$ 500 milhões no resultado anual das empresas. Assim aparece na figura 7.1 a 1ª. conclusão desta Dissertação: o processo proposto é uma solução de gerenciamento de receita que depende do tráfego de voz.

O PROCESSO PROPOSTO É UMA SOLUÇÃO DE GERÊNCIA DA RECEITA OPERACIONAL DE VOZ

Figura 7.1 1ª. Conclusão

O processo se inicia com a preparação de uma base de dados de grande abrangência cobrindo todas as áreas relevantes de geração de tráfego. As informações dos CDRs são recebidos em arquivos de dados gigantescos conforme apresentado no capítulo 1. Eles são tratados, filtrados, selecionados e formatados para gerar arquivos com

recebidos de uma dia de tráfego tem até 8 milhões de registros e o tamanho de até 350Mb. A proposta desta dissertação é utilizar ferramentas para capacitar um analista de tráfego, sem necessariamente ser um especialista em informática, lidar com este porte de arquivos. A idéia desta proposta é utilizar profissionais que tenham foco e conhecimento profundo em tráfego telefônico para poder realizar um gerenciamento completo do mesmo nas condições demandas: periodicamente e dentro de um prazo exíguo. A solução encontrada foi utilizar ferramenta de Business Inteligence como o SAS. O software foi utilizado na

-and- -and-

utilização pelo analista de tráfego e sem a mesma haveria a necessidade ter uma equipe de analistas e especialistas em informática dedicada e de prontidão. O software foi muito utilizado na etapa de preparação da base de dados onde foi necessário lidar com grandes

arquivos, mas mais ainda nas etapas de analise e atualização da base. Na análise do tráfego utilizaram-se muitos recursos de cálculo de parâmetros estatísticos, montagem de tabela e gráficos. Estes recursos viabilizaram a análise de arquivos gigantescos ao mesmo que proporcionaram mais flexibilidade. A agilidade e versatilidade para selecionar, filtrar e tratar as amostras e arquivos de tráfego é uma característica crucial para a etapa de atualização da base de dados, onde o trabalho deve ser realizado com extrema rapidez e como muita profundidade para descobrir comportamentos anômalos ou tendências emergentes e rápido para não perder o timing do operacional da empresa. A 2ª. Conclusão desta dissertação é apresentada na figura 7.2:
enviromen da ferramenta viabiliza o processo de gerenciamento do tráfego

A VERSÃO WINDOWING ENVIROMENT DO **SAS** VIABILIZA O PROCESSO DE GERENCIAMENTO TRÁFEGO PROPOSTO

Figura 7.2 2ª. Conclusão

Realizar previsões mensais a partir de um ambiente com uma base de dados completa devidamente atualizada e com uma ferramenta

disponíveis sobre o tráfego do mês. Teoricamente qualquer um sabe quantos dias tem um mês e é suficiente multiplicar os dias pelo valor representativo do tráfego disponível na base de dados conforme mostrado na tabela 6.5.

Na verdade, considerando a magnitude do tráfego envolvido de bilhões de minutos mês, o valor da receita operacional associada de bilhões de reais ano e o contexto do timming operacional da empresa com reuniões mensais e quinzenais para avaliar o resultado econômico-financeiro da empresa, deduz-se a necessidade de uma equipe profissional e dedicada. O projetista necessita ter conhecimento profundo de tráfego telefônico além de dominar o histórico. Não adianta ter uma base de dados abrangente, atualizada e com 2 ou 3 anos de histórico sem conhecê-la ou sem entender com ela foi formada. É importante ter informações sobre todos os eventos relevantes que possam impactar no comportamento do tráfego telefônico. É de conhecimento geral que o Brasil vai sediar em 2014 a Copa FIFA de futebol e o Rio de Janeiro uma Olimpíada em 2016. Como prever o impacto destes eventos? É necessário refletir profissionalmente sobre nisso, mas uma pista pode ser obtida analisando o impacto dos Jogos Pan-americanos de 2007 e das corridas de Formula 1 e Indy em São Paulo. Uma equipe de profissionais dedicada

poderá pesquisar o impacto da Copa FIFA de 2010 na Africa do Sul ou acompanhar as ações da British Telecom para a Olimpiada de 2012. A 3ª. Conclusão desta dissertação é mostrada na figura 7.3: somente uma equipe profissional e focada poderá maximizar os resultados obtidos no processo de gerenciamento do tráfego telefônico proposto nesta dissertação.

SOMENTE UMA EQUIPE PROFISSIONAL PODE MAXIMIZAR O RESULTADO DO PROCESSO DE GERENCIAMENTO TRÁFEGO PROPOSTO

Figura 7.3 3ª. Conclusão

7.1. RESUMO DA PROPOSTA DE PROCESSO

As operadoras Telecom que atuam no cenário nacional já perceberam que o nível de competição mercado brasileiro está crescendo muito, tanto para os grandes grupos de operadoras como Telefonica/Vivo, Embratel/Claro/Net, Oi e TIM/Intelig, para as médias como GVT/Vivendi, Nextel e CTBC/Algar e também para as pequenas empresas prestadoras de Serviço de Comunicação Multimídia-SCM. Quem quiser sobreviver neste ambiente deve ter conhecer bem o mercado e utilizar este conhecimento de uma forma ágil e ampla dentro da empresa. Um dos novos caminhos para este passo é a Analítica , nova ciência criada por Davenport (2007), cujo objetivo é capacitar a Empresa para a Competição Analítica que requer um processo de utilização massiva de dados e informações referentes ao negócio, usando análise quantitativa e estatística e também modelos exploratórios e de predição para orientar e direcionar as decisões e ações estratégicas das operadoras.

Nesta linha o objetivo desta dissertação é propor um processo de gerenciamento do tráfego telefônico eficaz e eficiente devido a sua importância e do seu impacto na receita operacional e conseqüentemente no resultado econômico das empresas de Telecom. O processo apresenta uma descrição de como obter uma base de dados abrangente tendo como exemplo a preparação de uma base para uma cidade com pouco mais de 1 milhões de habitantes e 42 centrais telefônicas. É apresentada também uma relação de dados que devem ser analisados e com uma demonstração do cálculo e definição da base dedados utilizando uma ferramenta de BI como o SAS para fazer uma série de análises no tráfego

da cidade tomada como exemplo. Em seguida é mostrada uma alternativa para atualizar e manter atualizada a base de dados utilizando ferramenta do SAS. Finalizando é apresentada uma alternativa para fazer a estimativa do tráfego onde são mostradas as informações mais relevantes a serem consideradas, apresentadas sugestões onde e como coletar estas informações, além de uma demonstração de como fazer a estimativa utilizando o mês de Fevereiro de 2010 como exemplo. A idéia geral desta dissertação é disponibilizar uma proposta de solução de Gerencia de Tráfego Telefônico de Voz completa, atualizada, ágil e confiável. A solução é completa porque descreve um processo desde a geração de uma base de dados até estimativa do tráfego mensal. É atualizada porque tem um processo específico permanente de atualização da base de dados e uma recomendação de buscar sempre informações de relevância que possam impactar no comportamento do tráfego. A solução é ágil porque utiliza uma ferramenta de BI com o SAS que facilita o acesso ao banco de dados, flexibiliza sua utilização por não especialistas em informática e agiliza e aprofunda as análises com a série de recursos estatísticos disponíveis. A solução é extremamente confiável porque é suportada por uma ferramenta robusta com boa aceitação no mercado e porque a recomendação de trabalhar com uma equipe profissional e dedicada é importante reforçar a recomendação de que na área de gerenciamento de tráfego não há espaço para improvisação ou amadorismo. Qualquer bobagem pode sair muito cara porque estamos lidando com bilhões de minutos associados a bilhões de reais.

7.2. FUTURAS SOLUÇÕES

Antes de concluir, é importante dar uma visão de que embora a solução apresentada seja completa, isto é, com inicio, meio e fim, ela não tem a pretensão de esgotar o assunto de análise e estimativa de tráfego. Existe um espaço enorme para novas análises. Lembrando de que no capítulo 1 foi mencionado que no CDR dispõe 69 campos de dados e nesta solução somente 6 dele106s foram utilizados. Quanta informação e quantas análises ainda podem ser feitas com os outros 63 campos? A demanda por mais informação e mais análises é percebida no dia-a-dia das empresas e podem ser relacionadas e classificadas vindo a se tornar objetos de outros trabalhos. A área de estimativa de tráfego é outro mundo aberto. A solução aqui apresentada foca no mês corrente mas as empresas tem outras demandas. As empresas publicam trimestralmente os

Balancetes contábeis[28] que são de extrema importância para os acionistas e investidores e que gera uma demanda de estimativa trimestral. Como estimar com antecedência o tráfego do trimestre? Todo ano as empresas montam o orçamento para o ano seguinte que é um dos itens mais importante de planejamento e controle. O orçamento anual define todos os objetivos da empresas para o período. Uma das receitas de maior relevância para a maioria das empresas de Telecom ainda é a receita operacional de voz. A previsão da receita de voz depende de uma boa estimativa do tráfego que somente poderá ser obtida através de um processo projeção de alta precisão.

Finalizando, é importante ressaltar que o processo proposto é mais que o gerenciamento de tráfego telefônico sendo na verdade uma solução de gerência da receita operacional de voz, viabilizada e suportada por uma ferramenta de BI como o SAS na versão

[28] Relatórios trimestrais com o resultado econômico financeiro das empresa amplamente divulgados

REFERÊNCIAS BIBLIOGRÁFICAS

ANATEL. Agência Nacional de Telecomunicações.Portal.*Resolução no. 432.* Brasília, Distrito Federal, Brasil, 2006

ANATEL. Agência Nacional de Telecomunicações.Portal.*Sistema de Documentação TELEBRÁS, Prática CPA-T, Requisitos Mínimos de Tarifação.* Brasília, Distrito Federal, Brasil, 1998

BORROR, C.M. ; Goldsman, D.M.; Hines,W.W., Montgomery, D.C. *Probabilidade e Estatística na Engenharia*, Rio de Janeiro, Rio de Janeiro, Brasil: LTC, 2006

CELEBRONI, Business Process Management. *Cadeia de Valor, Macro-Processo e Process*, 2009. Disponível em: <http:celebroni.blogspot.com> Acesso em 24 de Junho de 2011

COSTA, Silvio R.. *Análise Estatística na Conciliação da Receita de Público e Despesa de Uso de rede em Operadoras de Telecom.* Dissertação (Mestrado em Engenharia Elétrica). Universidade de Brasília, Brasília, Distrito Federal, Brasil, 2010

DAVENPORT, Thomas.; Harris, Jeanne. Competing on analytics: The new science of winning, Boston, Massachusets, USA: Harvard business school press, 2007

INFOESCOLA, Home Page do INFOESCOLA. Disponível em <http://www.infoescola.com.br>. Acesso em 14 de Maio de 2011

ITU-T, International Telecommunication Union Telecommunication Standardization Sector. *Recommendation Q.825, Specification of TMN applications at Q3 interface: Call detail recording*, 1998. Disponível em: <http://www.itu.int/> Acesso em: 15 de Maio 2011.

ITU-T, International Telecommunication Union Telecommunication Standardization Sector. *Recommendation X.730, Information Technology – open systems interconnection – systems management: object management function*, 1992. Disponível em: <http://www.itu.int/> Acesso em: 15 de Maio 2011

ITU-T, International Telecommunication Union Telecommunication Standardization Sector, 2001. *Recommendation E.502,Traffic engineering – Measurement and recording of traffic, Traffic measurement: requirements for digital telecommunication exchanges*. Disponível em: <http://www.itu.int/> Acesso em: 15 de Maio 2011

GLOBO, *BBB11 estatísticas*. Home Page do BBB11 em <http://bbb.globo.com>. Acesso em 14 de Maio de 2011

MCFEDRIES, P.*Fórmulas e Funções com Microsoft Office Excel 2007*. São Paulo, Brasil: Pearson Prentice Hall, 2009

OI, Relatório. *Relação com Investidores 1T2011*, 2011. Disponível em: <http://www.oi.com.brt/> Acesso em: 16 de Maio 2011

Odeh, R. e Owens, D. *Tables for Normal Tolerance Limits, Sampling Plans and Screening*. Reading, Massachusetts, USA: Addison Wesley Publishing Company, 1980

PORTAL ACTION, Home Page ACTION. Disponível em http://www.portalaction.com.br/. Acesso em: 10 de Maio de 2011

SAS, Home Page do SAS. Disponível em <http://www.sas.com.br>. Acesso em: 10 de Maio de 2011

SCHLOTZHAUER, S.D. *Elementary Statistics Using SAS*, Cary, NC, USA: SAS Institute Inc.Puttini. 2009

THE MATHFORUM. *Defining Quartiles*, Library, 2002. Disponível em: < http://mathforum.org/library/drmath/view/60969.html>. Acesso em: 16 de Maio de 2011

TELEFONICA, Relatório. *Resultados Janeiro – Março / 2011*. Disponível em: <http://www.telefonica.com.brt/> Acesso em: 16 de Maio 2011

TELETIME, *Atlas Brasileiro de Telecomunicações 2009*, São Paulo, SP: Converge Comunicações. 2009

TUDE, Eduardo. *Tráfego Telefônico (Erlang): Tutoriais Telefonia Fixa da Teleco*, 2003. Disponível em: <http://www.teleco.com.br>. Acesso em: 15 de Maio de 2011

ANEXOS

A – TESTE ESTATÍSTICO PARA NORMALIDADE DE SHAPIRO-WILK

Neste anexo é o descrito o teste para Normalidade de Shapiro-Wilk conforme PORTAL ACTION, (2011).

O teste Shapiro-Wilk é baseado na estatística W, calculada como a seguir:

$$W = \frac{b^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

onde x_i são os valores da amostra ordenados (x_1 é o menor). Menores valores de W são evidências de que os dados são normais. A constante b é determinada da seguinte forma

$$b = \sum_{i=1}^{n/2} a_{n-i+1} \times (x_{n-i+1} - x_i)$$

onde a_i são constantes geradas pelas médias, variâncias e covariâncias das estatísticas de ordem de uma amostra de tamanho n de uma distribuição Normal. Seus valores, tabelados, são dados abaixo:

$i \setminus N$	2	3	4	5	6	7	8	9	10	
1	0.7071	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.5888	0.5739	
2		0.0000	0.1677	0.2413	0.2806	0.3031	0.3164	0.3244	0.3291	
3				0.0000	0.0875	0.1401	0.1743	0.1976	0.2141	
4						0.0000	0.0561	0.0947	0.1224	
5								0.0000	0.0399	

$i \setminus N$	11	12	13	14	15	16	17	18	19	20
1	0.5601	0.5475	0.5359	0.5251	0.5150	0.5056	0.4968	0.4886	0.4808	0.4734
2	0.3315	0.3325	0.3325	0.3318	0.3306	0.3290	0.3273	0.3253	0.3232	0.3211
3	0.2260	0.2347	0.2412	0.2460	0.2495	0.2521	0.2540	0.2553	0.2561	0.2565
4	0.1429	0.1586	0.1707	0.1802	0.1878	0.1939	0.1988	0.2027	0.2059	0.2085
5	0.0695	0.0922	0.1099	0.1240	0.1353	0.1447	0.1524	0.1587	0.1641	0.1686
6	0.0000	0.0303	0.0539	0.0727	0.0880	0.1005	0.1109	0.1197	0.1271	0.1334
7			0.0000	0.0240	0.0433	0.0593	0.0725	0.0837	0.0932	0.1013
8					0.0000	0.0196	0.0359	0.0496	0.0612	0.0711
9							0.0000	0.0163	0.0303	0.0422
10									0.0000	0.0140

Para realizar o teste de Shapiro-Wilk, devemos:

1. Formulação da Hipótese:

$$\begin{cases} H_0 : \text{A amostra provém de uma população Normal} \\ H_1 : \text{A amostra não provém de uma população Normal} \end{cases}$$

2.

3. Calcular a estatística de teste:

   a. Ordenar as n observações da amostra: $x_1, x_2, x_3, \ldots, x_n$;

   b. Calcular $\sum_{i=1}^{n}(x_i - \bar{x})^2$;

   c. Calcular b;

   d. Calcular W;

4.

valores críticos da estatística W de Shapiro-Wilk são dados naTabela abaixo).

		Nível de significância α								
		0.01	0.02	0.05	0.10	0.50	0.90	0.95	0.98	0.99
Tamanho da Amostra, n	3	0.753	0.756	0.767	0.789	0.959	0.998	0.999	1.000	1.000
	4	0.687	0.707	0.748	0.792	0.935	0.987	0.992	0.996	0.997
	5	0.686	0.715	0.762	0.806	0.927	0.979	0.986	0.991	0.993
	6	0.713	0.743	0.788	0.826	0.927	0.974	0.981	0.986	0.989
	7	0.730	0.760	0.803	0.838	0.928	0.972	0.979	0.985	0.988
	8	0.749	0.778	0.818	0.851	0.932	0.972	0.978	0.984	0.987
	9	0.764	0.791	0.829	0.859	0.935	0.972	0.978	0.984	0.986
	10	0.781	0.806	0.842	0.869	0.938	0.972	0.978	0.983	0.986
	11	0.792	0.817	0.850	0.876	0.940	0.973	0.979	0.984	0.986
	12	0.805	0.828	0.859	0.883	0.943	0.973	0.979	0.984	0.986
	13	0.814	0.837	0.866	0.889	0.945	0.974	0.979	0.984	0.986
	14	0.825	0.846	0.874	0.895	0.947	0.975	0.980	0.984	0.986
	15	0.835	0.855	0.881	0.901	0.950	0.975	0.980	0.984	0.987
	16	0.844	0.863	0.887	0.906	0.952	0.976	0.981	0.985	0.987
	17	0.851	0.869	0.892	0.910	0.954	0.977	0.981	0.985	0.987
	18	0.858	0.874	0.897	0.914	0.956	0.978	0.982	0.986	0.988
	19	0.863	0.879	0.901	0.917	0.957	0.978	0.982	0.986	0.988
	20	0.868	0.884	0.905	0.920	0.959	0.979	0.983	0.986	0.988
	21	0.873	0.888	0.908	0.923	0.960	0.980	0.983	0.987	0.989
	22	0.878	0.892	0.911	0.926	0.961	0.980	0.984	0.987	0.989
	23	0.881	0.895	0.914	0.928	0.962	0.981	0.984	0.987	0.989
	24	0.884	0.898	0.916	0.930	0.963	0.981	0.984	0.987	0.989
	25	0.888	0.901	0.918	0.931	0.964	0.981	0.985	0.988	0.989
	26	0.891	0.904	0.920	0.933	0.965	0.982	0.985	0.988	0.989
	27	0.894	0.906	0.923	0.935	0.965	0.982	0.985	0.988	0.990
	28	0.896	0.908	0.924	0.936	0.966	0.982	0.985	0.988	0.990
	29	0.898	0.910	0.926	0.937	0.966	0.982	0.985	0.988	0.990
	30	0.900	0.912	0.927	0.939	0.967	0.983	0.985	0.988	0.900

B – TESTE ESTATÍSTICO PARA NORMALIDADE DE KOLGOMOROV-SMIRNOV

Neste anexo é o descrito o teste para Normalidade de Kolgomorov-Smirnov conforme o PORTAL ACTION, (2011).

O teste de Kolmogorov - Smirnov pode ser utilizado para avaliar as hipóteses:

$$\begin{cases} H_0 : \text{Os dados seguem uma distribuição normal} \\ H_1 : \text{Os dados não seguem uma distribuição normal.} \end{cases}$$

Este teste observa a máxima diferença absoluta entre a função de distribuição acumulada assumida para os dados, no caso a Normal, e a função de distribuição empírica dos dados. Como critério, comparamos esta diferença com um valor crítico, para um dado nível de significância.

A estatística utilizada para o teste é:

$$D_n = \sup_x |F(x) - F_n(x)|$$

Esta função corresponde a distância máxima vertical entre os gráficos de F(x) e $F_n(x)$ sobre a amplitude dos possíveis valores de x. Em D_n temos que:

- F(x) representa a função de distribuição acumulada assumida para os dados;
- $F_n(x)$ representa a função de distribuição acumulada empírica dos dados

Sejam $x_{(1)},\ldots,x_{(n)}$ observações aleatórias ordenadas de forma crescente da variável aleatória contínua X. A função de distribuição acumulada assumida para os dados é definida por $F(x_{(i)}$ $_{(i)})$ e a função de distribuição acumulada empírica é definida por uma função escada, dada pela fórmula:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{(-\infty,x]\}}(x_{(i)})$$

onde I_A é a função indicadora. A função indicadora é definida da seguinte forma:

$$I_A = \begin{cases} 1;\ \text{se } x \in A \\ 0;\ \text{caso contrário} \end{cases}$$

Observe que a função da distribuição empírica $F_n(x)$ corresponde à proporção de valores menores ou iguais a x. Tal função também pode ser escrita da seguinte forma

$$F_n(x) = \begin{cases} 0, \text{se } x<x_{(1)} \\ \frac{k}{n}, \text{se } x_{(k)} \leq x<x_{(k+1)} \\ 1, \text{se } x> x_{(n)} \end{cases} \quad (12)$$

Para realizar o teste, podemos considerar duas outras estatísticas:

$$D^+ = \sup_{x_{(i)}} |F(x_{(i)}) - F_n(x_{(i)})|$$
$$D^- = \sup_{x_{(i)}} |F(x_{(i)}) - F_n(x_{(i-1)})|$$

Essas estatísticas medem as distâncias (vertical) entre os gráficos das duas funções, teórica e empírica, nos pontos $x_{(i-1)}$ e $x_{(i)}$. Com isso, podemos utilizar como estatística de teste
$D_n = \max(D^+, D^-)$

Se D_n é maior que o valor crítico, rejeitamos a hipótese de normalidade dos dados com (1-

Resumo das estatísticas de teste.

x(ordenado)	$F_n(x$	$F(x) = P\left(z_{(i)} \leq \frac{x_{(i)} - \overline{x}}{s}\right)$	$\mid F(x_i) - F_n(x_{(i)}) \mid$	$\mid F(x_{(i)}) - F_n(x_{(i-1)})$
$x_{(1)}$	$\frac{1}{n}$	$F(x) = P\left(z_{(1)} \leq \frac{x_{(1)} - \overline{x}}{s}\right.$	$\mid F(x_{(1)}) - F_n(x_{(1)})$	$\mid F(x_{(1)}) - 0) \mid$
$x_{(2)}$	$\frac{2}{n}$	$F(x) = P\left(z_{(2)} \leq \frac{x_{(2)} - \overline{x}}{s}\right.$	$\mid F(x_{(2)}) - F_n(x_{(2)})$	$\mid F(x_{(2)}) - F_n(x_{(1)}) \mid$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\frac{n-1}{n}$	$F(x) = P\left(z_{(n)} \leq \frac{x_{(n)} - \overline{x}}{s}\right.$	$\mid F(x_{(n-1)}) - F_n x_{(n-}$	$\mid F(x_{(n-1)}) - F_n(x_{(n-}$
$x_{(n-1)}$				
$x_{(n)}$	1	$F(x) = P\left(z_{(n-1)} \leq \frac{x_{(n-1)}}{s}\right.$	$\mid F(x_{(n)}) - F_n(x_{(n)})$	$\mid F(x_{(n)}) - F_n(x_{(n-1)}$

Tabela : Estatísticas de teste.

OBS: O valor de $P\left(Z_{(i)} \leq \frac{x_{(i)} - \bar{x}}{s}\right)$ é encontrado na tabela da distribuição normal padrão.

A tabela de valores críticos para a estatística do teste de Komolgorov-Smirnov (D_n) é dada a seguir.

n	0,2	0,1	0,05	0,01
5	0,45	0,51	0,56	0,67
10	0,32	0,37	0,41	0,49
15	0,27	0,30	0,34	0,40
20	0,23	0,26	0,29	0,36
25	0,21	0,24	0,27	0,32
30	0,19	0,22	0,24	0,29
35	0,18	0,20	0,23	0,27
40	0,17	0,19	0,21	0,25
45	0,16	0,18	0,20	0,24
50	0,15	0,17	0,19	0,23
Valores maiores	$\frac{1,07}{\sqrt{n}}$	$\frac{1,22}{\sqrt{n}}$	$\frac{1,36}{\sqrt{n}}$	$\frac{1,63}{\sqrt{n}}$

C – TESTE ESTATÍSTICO PARA NORMALIDADE DE ANDERSON-DARLING

Neste anexo é o descrito o teste para Normalidade de Anderson - Darling conforme PORTAL ACTION, (2011).

O problema de inferência estatística que vamos considerar aqui é o de testar a hipótese de que uma dada amostra tenha sido retirada de uma dada população com função de distribuição acumulada contínua F(x), isto é, seja $x_1,x_2,...,x_n$ uma amostra aleatória e suponha que um provável candidato para a FDA dos dados seja F(x), então, o teste de hipóteses para verificar a adequabilidade da distribuição é:

$$\begin{cases} H_0: \text{ a amostra tem distribuição } F(x) \\ H_1: \text{ a amostra não tem distribuição } F(x) \end{cases}$$

Anderson e Darling (1952, 1954) propuseram a seguinte estatística para este teste

$$A^2 = n \int_{-\infty}^{\infty} \frac{[F_n(x) - F(x)]}{F(x)(1 - F(x))} dF(x)$$

onde $F_n(x)$ é a função de distribuição acumulada empírica definida como

$$F_n(x) = \begin{cases} 0, \text{se } x < x_{(1)} \\ \dfrac{k}{n}, \text{se } x_{(k)} \leq x < x_{(k+1)} \\ 1, \text{se } x > x_{(n)} \end{cases} \qquad (7.3.1)$$

e $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$, são as estatísticas de ordem da amostra aleatória.

A estatística A^2 pode ser colocada numa forma equivalente:

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} [(2i - 1) \ln(F(x_{(i)})) + (2(n - i) + 1) \ln(1 - F(x_{(i)}))]$$

A transformação $F(x_{(i)})$ leva $x_{(i)}$ em $U_{(i)}$ de uma amostra de tamanho n com distribuição uniforme em (0,1). Logo,

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} [(2i - 1) \ln(U_{(i)}) + (2(n - i) + 1) \ln(1 - U_{(i)})] \qquad (\star)$$

Para calcular o valor da estatística A^2 procedemos da seguinte forma:

- Ordenamos os valores da amostra: $x_{(1)}$ $_{(2)}$ $_{(n)}$;
- Quando necessário, estime os parâmetros da distribuição de interesse;
- Calcule $U_i=F(x_{(i)})$ e calcule o valor da estatística de Anderson Darling

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}[(2i-1)(\ln(U_i) + \ln(1-U_{n+1-i}))]$$

(na sua forma equivalente)

- Para cada uma das distribuições calcule, se for o caso, o valor da estatística modificada de acordo com as tabelas dadas para cada uma delas.

Para uma distribuição com parâmetros conhecidos temos os valores da função de distribuição acumulada da estatística A^2 tabulados em Peter and Lewis(1960). O problema surge quando um ou dois dos parâmetros da distribuição precisam ser estimados. Para contornar esse problema Stephens (1974, 1976, 1977) utilizou métodos assintóticos para tabular os valores dessas probabilidades quando os parâmetros das distribuições são desconhecidos.

Para a distribuição Normal com função densidade de probabilidade

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (-\infty<x<\infty).$$

A seguinte tabela fornece alguns valores de quantis e a estatística de Anderson Darling modificada.

Caso 0: ^{2}) é totalmente conhecido.

Caso 1: ^{2} é estimado por s^2.

Caso 2: ^{2} $\overline{x}$.

Caso 3: ^{2}) é conhecido e são estimados por $(\overline{x}, s^2)$

		Pontos percentis para cada				
Caso	**Modificação**	**15,0**	**10,0**	**5,0**	**2,5**	**1,0**
0	Nenhuma	1,610	1,933	2,492	3,070	3,857
1	-	0,784	0,897	1,088	1,281	1,541
2	-	1,443	1,761	2,315	2,890	3,682
3	$A^2(1+(4/n)-(25/n^2))$	0,560	0,632	0,751	0,870	1,029

D – PARÂMETROS ESTATÍSTICOS – DEFINIÇÕES

Neste anexo são descritos os parâmetros estatísticos utilizados pela ferramenta de BI utilizada nos cálculos e análises desta dissertação.

- Média Amostral

Considerando as observações x_1, x_2, ...,x_n , em uma amostra de tamanho *n* ,então a média amostral é:

$$\overline{X} = \frac{X_1 + \ldots + X_n}{n}$$

A média amostral X representa o valor médio de todas as observações da amostra.

- Desvio Padrão

  O desvio padrão amostral de um conjunto de dados é igual à raiz quadrada da variância amostral. Desta forma, o desvio padrão amostral é dado por:

$$s = \sqrt{s^2} = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1}}$$

- Variância amostral

  A variância de uma amostra $\{x_1,\ldots,x_n\}$ de n elementos é definida como a soma dos quadrados dos desvios de elementos em relação à sua média $\overline{x}$dividido por (n-1). Ou seja, a variância amostral é dada por:

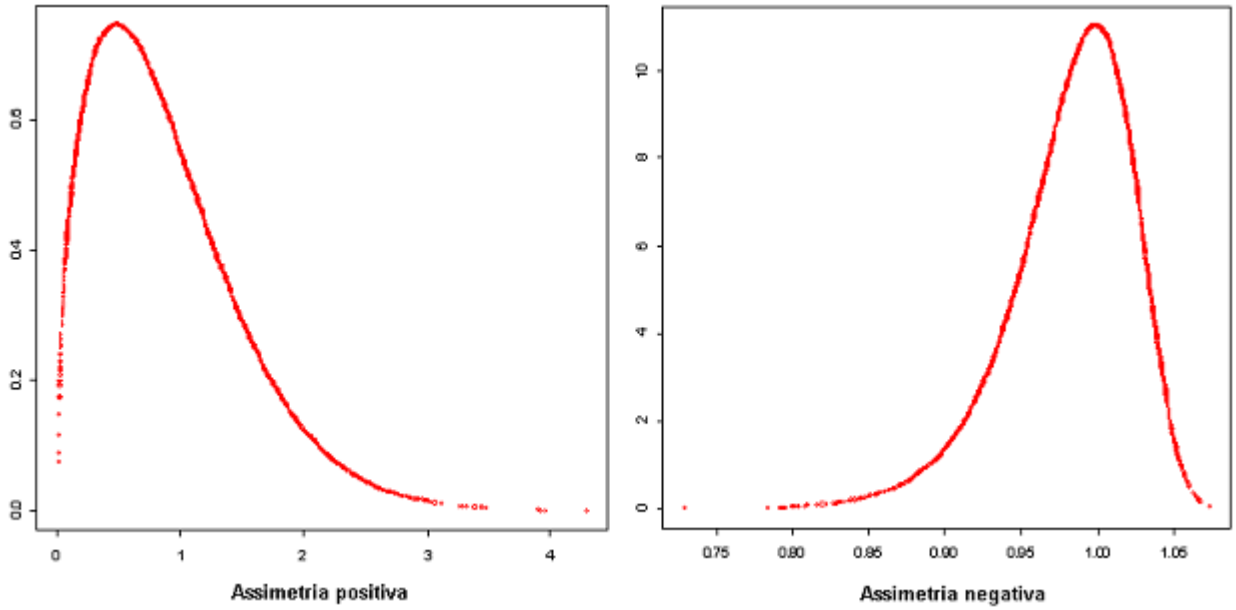
$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1}$$

  Ao utilizarmos a média amostral como estimador de m para calcularmos a variância amostral, perdemos 1 grau de liberdade em relação à variância populacional.

- Assimetria

  A assimetria permite distinguir as distribuições assimétricas. Um valor negativo indica que a cauda do lado esquerdo da função densidade de probabilidade é maior que a do lado direito. Um valor positivo para a

assimetria, indica que a cauda do lado direito é maior que a do lado esquerdo. Um valor nulo indica que os valores são distribuido de maneira relativamente igual em ambos os lados da média, mas não implica necessariamente, uma distribuição simétrica.

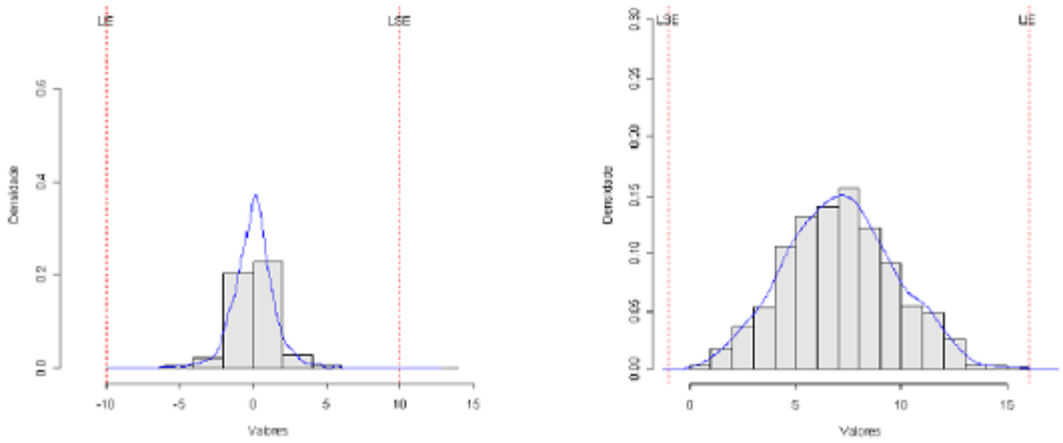


A fórmula da assimetria é dada por

$$b_1 = \frac{1}{n}\sum\left[\frac{x_i - \overline{x}}{s}\right]^3$$

- Curtose

A Curtose é uma medida de dispersão que caracteriza o "achatamento" da curva da função de distribuição. Um valor positivo costuma indicar um pico mais agudo, um corpo mais fino e uma cauda mais gorda que a distribuição normal. Um valor negativo indica um pico mais tênue, um corpo mais grosso e uma cauda mais fina que a distribuição normal. Para ilustrar a curtose, temos os gráficos a seguir, que representam uma curtose positiva (caudas mais alongadas) e curtose negativa (caudas mais curtas), respectivamente.



A fórmula da curtose é dada por

$$b_2 = \frac{1}{n}\sum\left[\frac{x_i - \overline{x}}{s}\right]^4 - 3$$

- Coeficiente de Variação

  O Coeficiente de Variação-CV segundo definição InfoEscola é a estatística utilizada quando se deseja comparar a variação de conjuntos de observações que diferem na média ou são medidos em grandezas diferentes (unidades de medição diferentes). O CV é o desvio padrão expresso como uma porcentagem média.

  CV = 100 . (s / Média) (%)

  O CV é uma medida relativa de variabilidade. É independente da unidade de medida utilizada, sendo que a unidade dos dados observados pode ser diferente que seu valor não será alterado.

  O coeficiente de variação tem, portanto, aplicações na pesquisa para comparar a precisão de diferentes experimentos. Entretanto, a qualificação de um coeficiente como alto ou baixo requer familiaridade com o material que é objeto de pesquisa.

- Erro Padrão da Média

  O erro padrão é uma medida da precisão da média amostral calculada. O erro padrão obtém-se dividindo o desvio padrão pela raíz quadrada do tamanho da amostra. Ou seja, $\frac{\sigma}{\sqrt{n}}$.

  Quando não se conhece o desvio padrão da população, usa-se o desvio padrão da amostra (**s**) ficando a formula:.

$$\frac{s}{\sqrt{n}}$$

Se de uma população, com média μ e desvio padrão σ, se retirarem muitas amostras todas do mesmo tamanho *n*, e para cada amostra se calcular a respectiva média, a distribuição de todas essas médias é normal com média μ e desvio padrão

$$\frac{\sigma}{\sqrt{n}}$$.

Assim, o erro padrão não é mais do que o desvio padrão da distribuição das médias das amostras de uma população

- Mediana

Para calcular a mediana devemos, em primeiro lugar, ordenar os dados do menor para o maior valor. Se o número de observações for ímpar, a mediana será a observação central. Se o número de observações for par, a mediana será a média aritmética das duas observações centrais. Notação: $\tilde{X}$.

- Quantis

Os quantis são pontos determinados em intervalos regulares de uma amostra de uma variável aleatória. Os quantis dividem o conjunto das observações da amostra em intervalos determinados de mesmo tamanho. Os percentis são quantis que dividem o conjunto de observações da amostra em 100. Os quartis são quantis que dividem o conjunto de observações da amostra em 4, sendo o primeiro quartil conhecido por Q1 ou quartil inferior, o segundo quartil conhecido por Q2 ou mediana e o terceiro quartil conhecido por Q3 ou quartil superior.

Segundo de THE MATHFORUM (2002) os:

> Quartiles are simple in concept but can be complicated in execution.
>
> The concept of quartiles is that you arrange the data in ascending order and divide it into four roughly equal parts. The upper quartile is the part containing the highest data values, the upper middle quartile is the part containing the next-highest data values, the lower quartile is the part containing the lowest data values, while the lower middle quartile is the part containing the next-lowest data values.

Here's where it starts to get confusing. The terms 'quartile', 'upper quartile' and 'lower quartile' each have two meanings. One definition refers to the subset of all data values in each of those parts. For example, if I say "my score was in the upper quartile on that math test", I mean that my score was one of the values in the upper quartile subset (i.e. the top 25% of all scores on that test).

But the terms can also refer to cut-off values between the subsets. The 'upper quartile' (sometimes labeled Q3 or UQ) can refer to a cut-off value between the upper quartile subset and the upper middle quartile subset. Similarly, the 'lower quartile' (sometimes labeled Q1 or LQ) can refer to a cut-off value between the lower quartile subset and the lower middle quartile subset.

The term 'quartiles' is sometimes used to collectively refer to these values plus the median (which is the cut-off value between the upper middle quartile subset and the lower middle quartile subset). John Tukey, the statistician who invented the box-and-whisker plot, referred to these cut-off values as 'hinges' to avoid confusion. Unfortunately, not everyone followed his lead on that.

It gets worse. Statisticians don't agree on whether the quartile values ('hinges') should be points from the data set itself, or whether they can fall between the points (as the median can when there are an even number of data points). Furthermore, if thequartile value is not required to be a point in the data set itself, most data sets don't have a unique set of values {Q1, Q2, Q3} that divides the data into four "roughly equal" portions. The SAS statistical software package, for example, allows you to choose from among five different methods for calculating the quartile values. How then do we choose the "best" value for the quartiles?

The answer to that question depends in part on the statisticians' objective in finding quartile values. Tukey wanted a method that was simple to use, "without the aid of calculating machinery." Others seek to minimize the bias in selecting the quartile values. Still others want methods that can be extended to other quantiles (for example, quintiles or percentiles). Thus, different methods have been developed for calculating the quartile values.

Tukey's method for finding the quartile values is to find the median of the data set, then find the median of the upper and lower halves of the data set. If there are an odd number of values in the data set, include the median value in both halves when finding the quartile values. For example, if we have the data set:

{1, 4, 9, 16, 25, 36, 49, 64, 81}

we first find the median value, which is 25. Since there are an odd number of values in the data set (9), we include the median in both halves. To find the quartile values, we must find the medians of:

{1, 4, 9, 16, 25} and {25, 36, 49, 64, 81}

Since each of these subsets has an odd number of elements (5), we use the middle value. Thus the lower quartile value is 9 and the upper quartile value is 49.

The TI-83 uses a method described by Moore and McCabe (sometimes referred to as "M-and-M") to find quartile values. Their method is similar to Tukey's, but you *don't* include the median in either half

when finding the quartile values. Using M-and-M on the data set above:

{1, 4, 9, 16, 25, 36, 49, 64, 81}

we first find that the median value is 25. This time we'll exclude the median from each half. To find the quartile values, we must find the medians of:

{1, 4, 9, 16} and {36, 49, 64, 81}

Since each of these data sets has an even number of elements (4), we average the middle two values. Thus the lower quartile value is (4+9)/2 = 6.5 and the upper quartile value is (49+64)/2 = 56.5.

With each of the above methods, the quartile values are always either one of the data points, or exactly half way between two data points.

Those methods involve only simple arithmetic and are easily extendable to octiles (eighths), hexadeciles (sixteenths), etc. They are not, however, extendable to quintiles (fifths) or percentiles (hundredths), etc. Furthermore, they tend to have a high bias. (That is, the quartile values calculated on subsets of the data set tend to vary more, and are not good predictors of the quartile values of the entire data set.)

Mendenhall and Sincich, in their text _Statistics for Engineering and the Sciences_, define a different method of finding quartile values. To apply their method on a data set with n elements, first calculate:

L = (1/4)(n+1)

and round to the nearest integer. If L falls halfway between two integers, round up. The Lth element is the lower quartile value. Next calculate:

U = (3/4)(n+1)

and round to the nearest integer. If U falls halfway between two integers, round down. The Uth element is the upper quartile value. So for our example data set:

{1, 4, 9, 16, 25, 36, 49, 64, 81} n = 9, so

L = (1/4)(9+1) = 2.5

which becomes 3 after rounding up. The lower quartile value is the 3rd data point, 9. Similarly:

U = (3/4)(9+1) = 7.5

which becomes 7 after rounding down. The upper quartile value is the 7th data point, 49.

Using this method, the upper and lower quartile values are always two of the data points.

Minitab uses the same method, except it doesn't round the values of L and U. Instead, it uses linear interpolation between the two closest data points. For our example above, instead of rounding L to

3, Minitab would let L = 2.5 and find the value half way between the 2nd and 3rd data points. In our example, that would be (4+9)/2 = 6.5. Similarly, the upper quartile value would be half way between the 7th and 8th data points, which would be (49+64)/2 = 56.5. If L were 2.25, Minitab would find the value one fourth of the way between the 2nd and 3rd data points and if L were 2.75, Minitab would find the value three fourths of the way between the 2nd and 3rd data points.

Excel uses a method described by Freund and Perles, which almost no one else uses. To apply this method on a data set with n elements, Excel first calculates L = (1/4)x(n+3). The Lth element is the lower quartile value. If L is not an integer, Excel uses linear interpolation. Next it calculates U = (1/4)x(3n+1). The Uth element is the upper quartile value. If U is not an integer, Excel again uses linear interpolation. So for our example data set:

{1, 4, 9, 16, 25, 36, 49, 64, 81} n = 9, so

L = (1/4)x(9+3) = 3

The lower quartile value is the 3rd data point, 9.

U = (1/4)x(3*9+1) = 7

The upper quartile value is the 7th data point, 49.

As we can see, these methods sometimes (but not always) produce the same results. To further illustrate, consider the following data sets:

A = {1, 2, 3, 4, 5, 6, 7, 8}
B = {1, 2, 3, 4, 5, 6, 7, 8, 9}
C = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
D = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11}

Here are the upper and lower quartile values, as calculated by each method described above:

	Tukey	M&M		M&S	Mini	Excel
Set A LQ:	2.5	2.5	2	2.25	2.75	
UQ:	6.5	6.5	7	6.75	6.25	
Set B LQ:	3.0	2.5	3	2.50	3.00	
UQ:	7.0	7.5	7	7.50	7.00	
Set C LQ:	3.0	3.0	3	2.75	3.25	
UQ:	8.0	8.0	8	8.25	7.75	
Set D LQ:	3.5	3.0	3	3.00	3.50	
UQ:	8.5	9.0	9	9.00	8.50	

E – ESTATÍSTICAS DAS OPERADORAS

O quadro deste anexo foi montado com informações colhidas em Teletime (2009), utilizando um percentual de participação da receita de voz sobre a receita líquida obtida por estimativa em Oi (2011) e Telefonica(2011).

	Operadora	Tipo	Receita Liq 3o. Trim 2008		Receita Anual Voz Ano		Part. Voz	Trimestre
1	Brasil Telecom	Fixa	R$	2,84	R$	4,55	40%	4
2	CTBC	Fixa	R$	0,32	R$	0,51	40%	4
3	Embratel	Fixa	R$	2,50	R$	4,01	40%	4
4	GVT	Fixa	R$	0,35	R$	0,56	40%	4
5	Intelig	Fixa	R$	-	R$	-	40%	4
6	Oi	Fixa	R$	4,75	R$	7,60	40%	4
7	Sercomtel	Fixa	R$	0,04	R$	0,06	40%	4
8	Telefonica	Fixa	R$	4,09	R$	6,55	40%	4
9	AEIOU	Móvel	R$	-	R$	-	50%	4
10	Brasil Telecom	Móvel	R$	0,48	R$	0,96	50%	4
11	Claro	Móvel	R$	2,95	R$	5,89	50%	4
12	CTBC	Móvel	R$	0,32	R$	0,64	50%	4
13	Oi	Móvel	R$	1,32	R$	2,64	50%	4
14	Sercomtel	Móvel	R$	0,01	R$	0,02	50%	4
15	TIM	Móvel	R$	3,36	R$	6,72	50%	4
16	Vivo	Móvel	R$	4,08	R$	8,16	50%	4
	Total		R$	27,40	R$	54,81	50%	4

No quadro acima, o campo Operadora mostra o nome de cada operadora. O campo Tipo informa se operadora é fixa ou móvel. O campo Receita Liq indica a receita líquida em R$ bilhões para 0 3º. Trimestre de 2008 onde para as operadoras inclui toda receita operacional e pata as operadoras móveis exclui a receita de Dados e Serviços de Valor Agregado-SVA. No campo Receita Anual, é calculada o valor anual da receita liquida. O campo Part. Voz traz o percentual da receita de voz para receita liquida.