APLICAÇÃO DE METODOLOGIA DE PRODUÇÃO DE CDR SINTÉTICO QUE PRESERVA A PRIVACIDADE DAS PESSOAS PARA O USO NA SIMULAÇÃO DE TRÁFEGO DE REDE DE TELEFONIA E ESTUDO DE MOBILIDADE URBANA

Francisco da Silva Medeiros

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia de Produção, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Produção.

Orientador: Marcos do Couto Bezerra Cavalcanti

Rio de Janeiro
Abril de 2016

APLICAÇÃO DE METODOLOGIA DE PRODUÇÃO DE CDR SINTÉTICO QUE PRESERVA A PRIVACIDADE DAS PESSOAS PARA O USO NA SIMULAÇÃO DE TRÁFEGO DE REDE DE TELEFONIA E ESTUDO DE MOBILIDADE URBANA

Francisco da Silva Medeiros

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE PRODUÇÃO.

Examinada por:

_____

Prof. Marcos do Couto Bezerra Cavalcanti, D.Sc.

_____

Prof. Marcus Vinicius de Araujo Fonseca, D.Sc.

_____

Prof. Liz Rejane Issberner, D.Sc.

RIO DE JANEIRO, RJ - BRASIL
ABRIL DE 2016

Medeiros, Francisco da Silva

Aplicação de metodologia de produção de CDR sintético que preserva a privacidade das pessoas para o uso na simulação de tráfego de rede de telefonia e estudo de mobilidade urbana/ Francisco da Silva Medeiros. – Rio de Janeiro: UFRJ/COPPE, 2016.

XI, 127 p.: il.; 29,7 cm.

Orientador: Marcos do Couto Bezerra Cavalcanti

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia de Produção, 2016.

Referências Bibliográficas: p. 100-103.

1. CDR Sintético. 2. Tráfego de rede de telefonia. 3. estudo de mobilidade urbana. I. Cavalcanti, Marcos do Couto Bezerra. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Produção. III. Título.

“*Como tudo está conectado a tudo e o que isso significa para os negócios, relações sociais e ciência*”
(Albert-László Barabási *em Linked*)

AGRADECIMENTOS

Agradeço aos meus pais, que fizeram muito no alcance do que eles podiam para minha formação.

Platão (2000) afirmou que o mais importante em tudo é o começo; Lakatos (1999) asseverou que acreditar na sua pesquisa científica é fundamental[1]. Ao meu orientador, professor Dr. Marcos do Couto Bezerra Cavalcanti, quem soube me ajudar ao longo do meu trajeto acadêmico. Agradeço ainda a todos os meus professores do curso de mestrado da Universidade Federal do Rio de Janeiro – UFRJ.

Ainda na área acadêmica, agradeço ao professor Demétrius de Souza, por me ajudar e estender seus conhecimentos ao longo dos meus estudos e suporte à minha dissertação.

No âmbito profissional, agradeço sinceramente todo o incentivo do amigo, do mestre e do professor Mauro Fukuda, diretor de engenharia e de tecnologia da empresa OI.

No âmbito pessoal agradeço aos meus amigos que revisaram meu trabalho acadêmico, complementando meus comentários, ideias e observações.

No âmbito familiar agradeço o suporte da minha esposa e do meu filho que me ajudaram, me confortaram em momentos bem difíceis, impulsionando-me até ao fim deste desafio.

[1] Lakatos (1999): “A theory may even be of supreme scientific value even if no one understands it, let alone believes it.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

APLICAÇÃO DE METODOLOGIA DE PRODUÇÃO DE CDR SINTÉTICO QUE PRESERVA A PRIVACIDADE DAS PESSOAS PARA O USO NA SIMULAÇÃO DE TRÁFEGO DE REDE DE TELEFONIA E ESTUDO DE MOBILIDADE URBANA

Francisco da Silva Medeiros

Abril/2016

Orientador: Marcos do Couto Bezerra Cavalcanti

Programa: Engenharia de Produção

Esta dissertação tem por objetivo apresentar didaticamente a implementação da técnica de geração de *Call Detail Record* (CDR) "sintético" (hipotético), que amplia o uso de CDR em pesquisas acadêmicas e estudos empresariais, garantindo a privacidade das pessoas. Através do método ***W**ork and **H**ome **E**xtracted **RE**gions* (WHERE*)*, que gera o CDR sintético, é possível simular o tráfego de rede de telefonia e padrões de mobilidade urbana. A possibilidade de se estudar o tráfego de rede prevendo ações para aumentar a qualidade, disponibilidade e extensão da rede móvel de telecomunicações, além da alternativa de prever o deslocamento massivo urbano de pessoas, posiciona esta dissertação como ferramenta para a melhoria de serviços de transportes urbanos. Para não fazer uso de CDRs reais na construção de CDRs sintéticos, acrescentou-se ao método WHERE dois trabalhos matemáticos que contribuíram para tornar o resultado desta dissertação original: (i) a construção de CDRs sintéticos com o mesmo padrão original, reproduzindo a mesma variação temporal de frequência das ligações (ii) o sorteio dos horários das ligações de cada usuário, de acordo com curva de probabilidade que define o padrão de ligação de uma determinada população, além de contornar com dados públicos todas as dificuldades práticas que nascem da falta de dados reais de CDR. Assim, é possível criar cenários que possam inferir sobre a mecânica do deslocamento das pessoas em grandes cidades, contribuindo para a melhoria de políticas públicas integradas ao desenvolvimento regional e ao avanço tecnológico do país.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

APPLICATION OF A SYNTHETIC CDR PRODUCTION METHOD TO PRESERVE THE PRIVACY OF PEOPLE FOR USE IN TELEPHONE NETWORK TRAFFIC SIMULATION AND FOR THE STUDY OF URBAN MOBILITY

Francisco da Silva Medeiros

April/2016

Advisor: Marcos do Couto Bezerra Cavalcanti

Department: Production Engineering

This paper aims to present didactically the implementation of a technique for the production of “synthetic” (hypothetical) Call Detailed Records (CDRs) which expands the use of CDRs in academic research and corporate analysis whilst ensuring the privacy of users. Through the Work and Home Extracted Regions method (WHERE), which generates the synthetic CDR, one can simulate telephone network traffic and urban mobility patterns. The possibility of studying network traffic forecasting actions to increase the quality, availability and reach of the physical network of mobile access telecommunications - as well as the alternative of studying and predicting the massive urban displacement of people - positions this dissertation as a fundamental tool for the improvement of urban transport services. In order to avoid the use of real CDRs in the construction of synthetic CDRs, this dissertation added to the WHERE method two mathematical processes that help to make this dissertation original: (i) the construction of synthetic CDRs with the same formatting pattern as the original CDRs, thus reproducing the same temporal variation frequency of the original CDRs connections; (ii) randomly selecting the times of the calls of each user, according to the probability curve that defines the pattern of calls of a certain population, thus avoiding using public data to get around all the practical difficulties that result from the lack of real CDR data. Thus, one can create hypothetical scenarios that can make inferences about the mechanics of the displacement of people in large population centers, contributing to the improvement of public policies integrated with regional development and technological advances within the country.

SUMÁRIO

ÍNDICE DE FIGURAS

ÍNDICE DE QUADROS

ÍNDICE DE TABELAS

ÍNDICE DE GRÁFICOS

1 INTRODUÇÃO

Um dos principais marcos para a ciência brasileira foi o desembarque de Oswaldo Cruz no Brasil após seus estudos na França. Esse evento simboliza um divisor de águas no desenvolvimento da produção de conhecimento científico por brasileiros, em solo brasileiro. Em sua obra, Cukierman (2007) descreve a posição do médico em defender a importância de fortalecer, desenvolver e valorizar o conhecimento e a cultura local brasileira. Segundo o autor, Cruz alertava que o país não poderia se limitar a repetir as teorias já prontas, estudadas e escritas no exterior. O ponto de vista defendido por Oswaldo Cruz era de que a ciência é a base do progresso do conhecimento, proporcionando o desenvolvimento intelectual do homem e, consequentemente, a mudança de padrões sociais.

Paralelamente, Stewart (1998, p. XVII) afirma que "o conhecimento tornou-se o fator mais importante da produção" e principal elemento gerador de riqueza. Nesse contexto,"transformar dados em informação e em conhecimento, gerando ideias as quais, na verdade, são procedimentos que usamos para reconfigurar coisas físicas que já existem" (FONSECA, 2013, p. 32).

Atualmente as pessoas utilizam seus telefones móveis para realizarem ligações telefônicas constantemente. Esse comportamento revelou uma nova identidade de perfil humano – a figura do "ligador". Esse movimento vem ocorrendo, em função do espaço que a telefonia móvel passou a ocupar na sociedade contemporânea. Em consequência, o aparelho celular se tornou o mais importante sensor de presença existente em nossa época. Hoje, segundo estudos anteriores, as pessoas carregam seus aparelhos celulares para todos os lugares que elas frequentam. Essa característica permitiu que o estudo sobre a mobilidade urbana – ou melhor dizendo – estudo sobre a localização geográfica das pessoas se tornasse mais assertivo e revelasse onde elas mais frequentam, como a localização física do trabalho, da residência e dos lugares mais frequentados no lazer, ou seja, seu deslocamento pela cidade diariamente. As pessoas ligam, acessam a internet ou enviam mensagem de texto de qualquer lugar, a qualquer momento do dia. Esse novo hábito registra nas centrais telefônicas das operadoras de telefonia, em tempo real, esse movimento através dos registros de ligações telefônicas – chamados de CDR. Os CDR são gerados automaticamente pelas centrais telefônicas, toda vez que é realizada uma ligação telefônica de voz, acesso de dados ou mensagens sejam em rede fixa, rede

móvel ou redes IP. Os CDR revelam a identidades das pessoas melhor até que a impressão digital, segundo estudos.

A telefonia celular modificou o cotidiano das pessoas radicalmente. De fato, o aparelho celular tornou-se um dispositivo essencial e de valor intangível para as pessoas. A tecnologia de telefonia móvel mantém-se em constante evolução, com grandes desafios para cobrir países de grande expansão territorial e gerando muitas oportunidades para o consumidor e para a indústria das telecomunicações. Hoje, por meio da tecnologia móvel, é possível localizar as pessoas; não só onde elas estão, mas também de onde vieram. Um dos métodos utilizados para identificar esse movimento é através do *Call Detail Record* (CDR), que são gerados, automaticamente, pelos elementos de redes das operadoras de telecomunicações quando da realização de uma chamada (com ou sem sucesso). Segundo Isaacman (2012)[2], os dispositivos móveis dão a oportunidade aos pesquisadores de compreenderem como e quando as pessoas se movem, bastando apenas que realizem uma ligação telefônica. Por este motivo, o estudo de deslocamento urbano das pessoas vem se tornando cada vez mais importante no suporte ao desenvolvimento das cidades.

Segundo pesquisa sobre mobilidade urbana, realizada pelo Instituto de Pesquisa Econômica Aplicada (IPEA, 2011)[3], o conceito de mobilidade é entendido como a facilidade de deslocamento; por vezes, vincula-se àqueles que são transportados ou se transportam e, por outras vezes, relaciona-se à cidade ou ao local onde o deslocamento pode acontecer. Segundo o Instituto Brasileiro de Geografia e Estatística (IBGE, 2011), a mobilidade pendular[4] é caracterizada pelo deslocamento da população no território e num contexto determinado, por exemplo, da relação entre o domicílio e o trabalho ou vice-versa. Neste sentido, a mobilidade pendular contemporânea requer novos estudos que abordem o deslocamento de pessoas em sua vida cotidiana e que sejam levados em consideração, elementos como distância, duração, frequência, retenção, situação político-administrativa, redes sociais e urbanas, formas de deslocamento e motivações para as pessoas mudarem-se de lugar. Ainda, segundo o IBGE (2011), a mobilidade

[2] Isaacman (2012, p. iii): *...mobile devices gives researchers a chance to understand how and when people move...*

[3] Pesquisa, no tema mobilidade urbana, foi realizada por meio de entrevistas domiciliares, num total de 2.786 questionários válidos (com 30 questões) aplicados a pessoas maiores de 18 anos entre os dias 4 e 20 de agosto de 2010 em 146 municípios. Considerou-se uma distribuição pelas grandes regiões do país e por cotas, tendo como parâmetros a Pesquisa Nacional por Amostragem de Domicílios (PNAD) 2008, realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE).

[4] Mobilidade pendular: refere-se ao deslocamento das pessoas do domicilio (lugar de origem) e o trabalho (lugar de destino).

pendular associa-se à questão da infraestrutura urbana, especialmente em relação aos transportes urbanos municipal e intermunicipal, e possibilita ampliar a questão da permanência no local de trabalho e/ou estudo. Isso porque o estudo sobre mobilidade pendular indaga, pela primeira vez, se as pessoas retornam do trabalho para casa diariamente e qual é o tempo habitual gasto no deslocamento de sua casa até o trabalho. Entretanto, esta dissertação está longe de apontar soluções para a questão dos transportes urbanos, mas introduzir o método ***W**ork and **H**ome **E**xtracted **RE**gions* (WHERE) para a produção de CDR sintético, que permite realizar pesquisas com CDR que contribuem para o estudo do deslocamento de pessoas pela cidade e, também, para a simulação de tráfego de rede de telefonia para planejamento, expansão e implantação de infraestrutura da rede física de telecomunicações.

Por esse motivo, este estudo é o resultado de uma proposta de mudança a partir de uma intervenção, o que o torna um problema de engenharia. Para uma situação ser caracterizada como um problema de engenharia, é necessário que sejam identificadas quatro[5] características (KOEN, 2003, p.11-24)[6]: (i) mudanças – a serem geradas; (ii) recurso – requer diferente uso para diferente material utilizado; (iii) solução – buscar sempre a melhor alternativa para a sociedade; e, por fim, (iv) a incerteza – se a técnica aplicada, de fato, solucionará o problema. Nesta pesquisa, o resultado final dos dados captados[6] é criar novos serviços utilizando CDR como fonte de dados para direcionar a inovação. Por meio da utilização do método WHERE, é possível criar cenários que podem inferir sobre a mecânica da mobilidade urbana em grandes centros urbanos, o que pode contribuir para a melhoria de políticas públicas, de forma integrada com o desenvolvimento regional e o avanço tecnológico. Essa proposição de intervenção implicará criar novos processos de inovação nas empresas, novas formas de relacionamentos entre operadoras de telecomunicações e seus usuários, de modo que seja possível melhor atender às necessidades da população e da oferta de novos serviços. Essa habilidade para resolver problemas complexos ou, ainda, reduzir a insatisfação dos usuários é o que identifica a terceira característica da essência da solução de um problema de engenharia – a heurística[7] (KOEN, 2003).

[5] Koen (2003, p.11): Now we will look in detail at the key words change, resources, best, and uncertainty, all of which have appeared in the definition of an engineering problem situation.
[6] Refere-se à fase de obtenção e identificação dos dados levantados nos testes.
[7] Koen (2003, p.28): Engineering design is the essence of engineering" and "A heuristic is anything that provides a plausible aid or direction in the solution of a problem but is in the final analysis unjustified, incapable of justification, and potentially fallible.

Por esse motivo, a importância da habilidade em resolver problemas torna-se mais necessária ao considerar o crescimento exponencial de dados gerados a todo segundo, por diversas bases e dispositivos conectados à internet. Segundo o Portal de informações do setor de telecomunicações do Brasil (TELECO, 2015), os números do final do ano de 2014 de telefones móveis alcançaram 280,7 milhões, os de telefones fixos chegaram a 45 milhões, os números de banda larga atingiram 24 milhões, os de TV por assinatura, 19,6 milhões, e as conexões máquina-máquina (M2M) chegaram a 10 milhões. Dessa forma, é essencial às empresas de telecomunicações selecionarem que informações são mais significativas, a fim de criar serviços mais inteligentes e analisarem esses dados em grande escala, com múltiplas variedades e em alta velocidade, para gerarem informações relevantes.

O recorte escolhido dentro do segmento de telecomunicações foi o registro de ligações telefônicas (CDR). Aginsky (2012)[8] ressalta que a indústria das comunicações, em particular as operadoras, apresentam condições mais favoráveis para esse estudo, tendo em vista o volume, a variedade de dados e a velocidade de processamento desses dados em tempo real (3Vs)[9].

1.1 JUSTIFICATIVAS PARA A ESCOLHA DO TEMA

A facilidade da indústria de telecomunicações em oferecer cenários mais propícios para o estudo de CDR e a experiência do pesquisador nos últimos 13 anos de um trajeto de carreira profissional desenvolvida dentro de um centro de pesquisa, de desenvolvimento e de testes de uma das grandes empresas de telecomunicações do mercado brasileiro motivou o tema desta dissertação. Essa experiência profissional foi fortemente favorecida pelos seguintes fatores: (i) a troca de experiências técnicas com os fabricantes de dispositivos de telecomunicações; (ii) o acesso às insatisfações dos usuários dos serviços neste segmento; e (iii) a discussão técnica em Comunidades de Prática (CP)[10], que são definidas como "oficinas do capital humano" e o "lugar onde as coisas acontecem" (FONSECA, 2013, p. 25).

[8] Aginsky (2012): The communications industry was made for 'Big Data' - managing vast volumes in real-time.

[9] Aginsky (2012): Big Data is characterized by three V's: Volumes (demand for voluminous data), Variety (new types, varied and unstructured, from new sources), and Velocity (rapidly analyzed, through distributed processing).

[10] São formadas por "equipes de profissionais informalmente induzidos a somar *expertises* e com predileção por um empreendimento conjunto. Elas são de natureza espontânea, orgânica e informal,

Os celulares estão, geralmente, muito próximos das pessoas. Elas carregam o aparelho de um lado para o outro e, cada vez mais, esse dispositivo ganha novas funcionalidades, tornando-o indispensável. Este comportamento reforça o fato de que os celulares sejam utilizados como sensores, e são excelentes para esse objetivo, já que estão sempre próximos das pessoas. A partir deste conceito, o aparelho móvel tornou-se um dispositivo importantíssimo para o estudo da mobilidade das pessoas, identificando onde estavam e para onde se deslocaram. Através de uma ligação ou de uma mensagem de texto enviada, é possível identificar a localização do indivíduo, o que serve para saber, em grandes eventos (jogos de futebol, *shows*, comícios etc.), o número aproximado de pessoas presentes no local. Além disso, através do CDR, é possível identificar os hábitos de cada um e seu comportamento, mediante o uso do aparelho móvel no seu dia a dia. Esta dissertação tem como objetivo principal demonstrar, passo a passo, a técnica de produção de CDR sintético, garantindo a privacidade das pessoas, aplicada ao estudo de mobilidade urbana e contribuindo no planejamento de rede de telefonia móveis.

Diante dessas influências, a busca por mais conhecimento sobre o assunto CDR foi natural. A relevância[11] deste estudo consiste em trabalhar ou manipular os dados de CDR, com toda a riqueza de informação, sem violar a privacidade das pessoas, pois, neste trabalho, a geração de CDR sintético não faz uso de CDRs reais na construção do CDR sintético.

A importância desta dissertação, para a ciência, é (i) apresentar uma técnica de geração de CDR, tal que os pesquisadores possam ampliar seus estudos sobre mobilidade urbana, sem ter que manipular informações privadas que, hoje, estão disponíveis nos CDRs reais, sob responsabilidade das operadoras de telecomunicações. Para as operadoras de telecomunicações, (ii) a técnica de criação de CDR sintético permitirá planejar melhor a implantação e a expansão de redes de telefonia móvel e de rede fixa, pois, através de CDR sintético, é possível criar populações "sintéticas", cidades "sintéticas" (populações e cidades hipotéticas) e introduzir pertubações nesses cenários para observar novos ou, ainda, comportamentos não previsíveis em cenários reais de redes de telefonia. Para a sociedade, (iii) o benefício está na melhoria de qualidade das redes e dos serviços de telecomunicações, principalmente em países como

tornando-se resistente à supervisão e interferência. Cooperam entre si, sondam-se mutuamente, ensinam umas às outras, exploram juntas um novo assunto" (FONSECA, 2013, p. 24).

[11] Relevância, neste caso, entende-se como o problema da dissertação. A técnica utilizada para determinar o problema da dissertação é descrita na obra de BOOTH et al. (2008).

o Brasil, onde o espaço territorial é grande. Por todas essas diferentes oportunidades, o estudo de CDR é fundamental para a indústria das telecomunicações.

1.2 CONTEXTUALIZAÇÃO DO TEMA

Após o processo de privatização dos serviços de telecomunicações no Brasil, as operadoras passaram a buscar novos serviços que as diferenciassem das concorrentes e, assim, aumentassem a abrangência de seus mercados. É importante ressaltar que o papel da agência reguladora do governo – Agência Nacional de Telecomunicações (ANATEL) – foi essencial para assegurar, de forma transparente, as importantes regulamentações que modificaram, de maneira radical, os diversos serviços de telecomunicações aos quais atualmente se tem acesso.

A ANATEL, como órgão oficial regulador das operadoras de telecomunicações, tem como uma de suas atribuições monitorar e acompanhar os indicadores de qualidade dos serviços de redes de telefonia prestados pelas operadoras no Brasil. No Gráfico 1 a seguir, pode-se observar que uma das maiores preocupações da agência está na qualidade da rede de telefonia móvel.

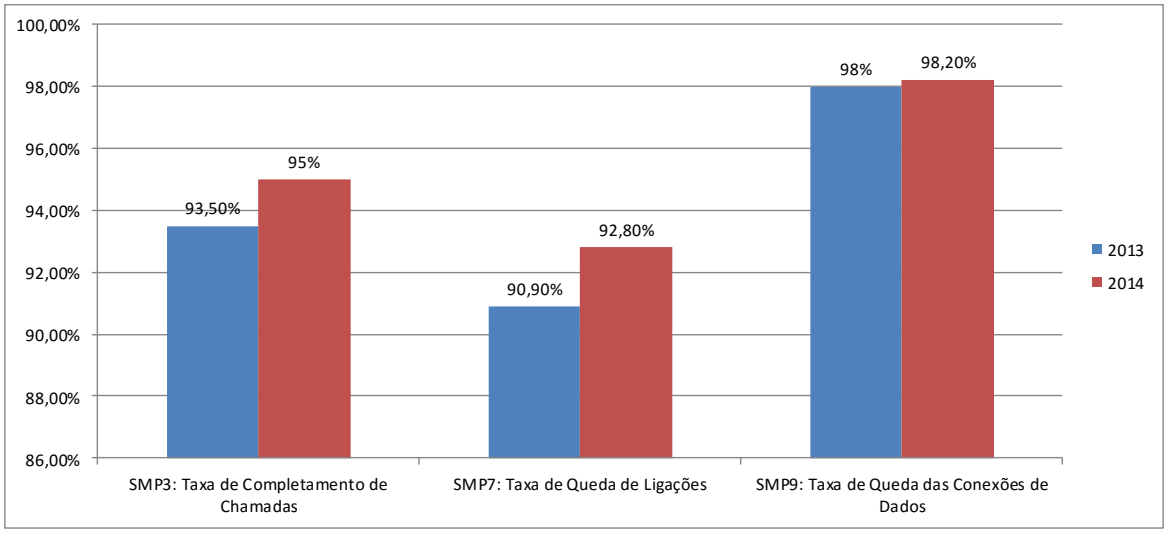


Gráfico 1 - Percentual de Indicadores com Cumprimento de Metas
Fonte: ANATEL (2014)

A situação é crítica, especialmente, na área de telefonia celular, caracterizada pelo sinal ruim e pelos altos índices de insatisfação registrados nos órgãos de defesa do consumidor, em relação a um serviço que não é barato.

O primeiro indicador de Serviço Móvel Pessoal (SMP), o SMP3, é a razão entre o total de chamadas originadas na rede da prestadora e o total de chamadas atendidas nos sistemas de autoatendimento do usuário (o telefone móvel). O segundo indicador, SMP7, é a razão entre o total de chamadas interrompidas por queda de ligação e o total de chamadas completadas com sucesso. O terceiro indicador, SMP9, é a razão entre o total de quedas de conexões de dados e o total de tentativas de conexão. Todos esses três indicadores são medidos nos Períodos de Maior Movimento (PMM), estabelecidos pela ANATEL.

Outros estudos realizados pelo mercado de telecomunicações apresentam dados alarmantes na qualidade de rede das operadoras de telefonia, que afetam, sensivelmente, a satisfação de seus clientes. Segundo estudo realizado pela associação TeleManagement Forum (TMForum, 2011), 30% dos problemas na operação de uma operadora está no provisionamento de serviços de rede, 28% está relacionado à qualidade da rede de telefonia disponibilizada aos clientes e 15% está associada à indisponibilidade de serviço. Assim, o percentual de 73% são de problemas de planejamento de rede de telefonia celular.

A maior parte desses problemas são identificados através de desconexões de ligações de telefones e o motivo deste comportamento é observado no CDR.

Diante de tantos desafios, o recorte escolhido e desenvolvido neste projeto foi o CDR, que é o registro de uma ligação telefônica, a qual pode ser realizada utilizando-se diversas tecnologias de rede, tal como a fixa, móvel, IP e heterogênea. Desta forma, pode-se, seguramente, afirmar que toda e qualquer chamada telefônica possui o registro da ligação efetuada por um terminal – fixo ou móvel – ou por outros equipamentos, como computadores.

O tamanho da base de dados do CDR é da ordem de grandeza de 1 petabyte. Considerando que cada terminal possui um registro de CDR, o volume de dados a ser armazenado, diariamente, é bastante relevante. O CDR é composto por diferentes campos de dados como: hora inicial, hora final, duração, número de quem efetuou a ligação e de quem a recebeu, se a ligação foi completada com sucesso ou se apresentou erro, mensagem de sinalização entre centrais telefônicas ou elementos de redes, unidade federativa, identificação da prestadora do serviço de telecomunicação, entre muitos outros campos. Por esse motivo, o tamanho dessa base é grande, ou melhor, é gigantesca.

Segundo o relatório *Business opportunities: Big Data*, da comissão europeia de emprego, crescimento e investimento da União Europeia (UE, 2013)[12], as empresas de telecomunicações estão entre as precursoras para a adoção da plataforma de *Big Data*, impulsionada por aplicações de dados intensivos, tais como registros de dados de chamadas (CDRs), monitoramento de tráfego de rede e de conteúdo digital. Ainda, segundo o relatório da comissão europeia, o termo *Big Data* descreve o aumento contínuo de dados e as tecnologias necessárias para coletar, amarzenar, gerenciar e analisar estes dados.

Segundo Hu, Wen e Chua (2014), o *Big Data* requer grande volume de dados, conforme ilustrado na Figura 1 a seguir.

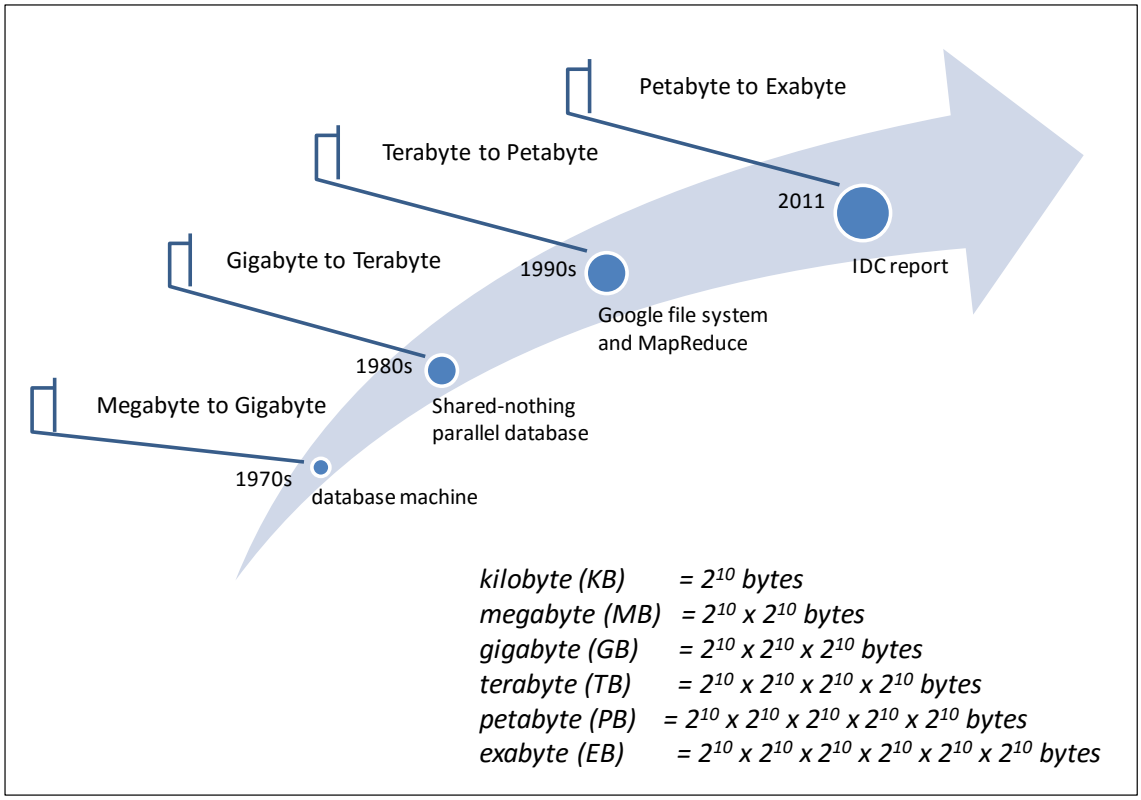


Figura 1 - Principais marcos da Plataforma de *Big Data*
Fonte: Hu et al., 2014

[12] A União Europeia (UE) representa os interesses dos países do mercado comum europeu e propõe nova legislação ao Parlamento Europeu e ao Conselho da União Europeia e assegura a correta aplicação do direito europeu pelos países da UE.

A Figura 1[13] apresenta os principais marcos da evolução e do crescimento do tamanho da base de dados, que se inicia em Megabyte e termina em Exabyte.

O CDR, ao longo dos últimos anos, foi se consolidando como uma fonte preciosa de exploração de dados. Os telefones móveis criam oceanos de informações sobre onde os indivíduos estão, para onde irão, o que eles estão comprando. Combinando tudo isso com outras informações, como dados da economia, da política, de redes sociais e muito mais, será possível abrir novos mundos (WEINBERGER, 2011). Atualmente, a aplicabilidade dessa base é estudada, de forma isolada, por fabricantes, operadoras de telecomunicações, pesquisadores e órgãos de legislação, de normatização e de pesquisa dos governos Federais, Estaduais e Municipais. Para manipular a base de CDR, é necessária prévia autorização das operadoras de telefonia, pois são quem têm os direitos legais sobre estas informações. Contudo, há o risco de, na manipulação dos dados de CDR, que a privacidade da identidade das pessoas (que realizaram e receberam as ligações) seja quebrada, o que pode gerar diversos inconvenientes legais sobre a pesquisa. Segundo relatório de pesquisadores do MIT e da Universidade de Louvain, na Bélgica, 95% dos usuários de telefones celulares, mesmo anonimizados, podem ser identificados com base em padrões de seus movimentos (MONTJOYE at al., 2013). Para Clarke (1999), o conceito de anonimização está relacionado à remoção das informações de identificação pessoal de cada registro. Por este motivo, os pesquisadores, atualmente, substituem os números de telefones das pessoas dos campos dos CDRs por valores numéricos sequenciais.

Os telefones móveis são registrados pelas antenas de telefonia quando se deslocam de um lugar para o outro, ou mesmo quando não estão sendo utilizados para algum serviço. Estes aparelhos móveis podem ser utilizados para a localização das pessoas ao longo do seu deslocamento dentro das cidades. Por meio de técnicas de *Big Data*, de análise de dados de grandes volumes de dados (1,5 milhão), de registros de CDRs móveis anônimos de uma grande operadora de celular da Europa, ao longo de 15 meses de pesquisa, foi possível identificar as pessoas individualmente. Os registos de localização de usuários anônimos apresentaram movimentos de padrões de comportamento idênticos, que eram tão únicos como impressões digitais. Os pesquisadores concluíram, com este estudo, que a preservação de dados anônimos não

[13] Hu et al. (2014, p. 655): "A brief history of big data with major milestones. It can be roughly split into four stages according to the data size growth of order, including Megabyte to Gigabyte, Gigabyte to Terabyte, Terabyte to Petabyte, and Petabyte to Exabyte".

é, necessariamente, suficiente para garantir a privacidade real (MONTJOYE et al., 2013).

Outro importante estudo realizado por Zang e Bolot (2011) foi responder ao seguinte questionamento: Deve-se ou não confiar na privacidade de publicação dos dados anonimizados de CDR? A constatação final foi de que a técnica de anonimização, de fato, não funciona. A pesquisa contemplou 30 bilhões de registros telefônicos de chamada de voz realizadas por 25 milhões de usuários de telefones móveis em todos os 50 estados dos EUA, no período de três meses, em uma operadora de telecomunicações de abrangência nacional de telefonia. Esse trabalho foi um divisor de águas, pois nenhum estudo dessa magnitude havia sido realizado.

É indiscutível que os registros de ligações telefônicas são ricas em informações. Os CDR impulsionaram o estudo sobre o deslocamento urbano das pessoas devido a dois principais fatores: (i) identificar a localização geográfica das pessoas, onde elas estão realizando suas ligações telefônicas e (ii) identificar o horário exato que que as ligações foram realizadas, podendo assim inferir sobre o tempo de deslocamento das pessoas ao longo da cidade. Toda essa informação expõe o cotidiano das pessoas publicamente, pois é possível inferir o local de residência, o local de trabalho e os locais que as pessoas mais frequentam. Além disso, o CDR contém os telefones das pessoas, o que por lei é informação restrita e de domínio privado. Por este motivo, os registros de ligações telefônicas são de propriedade exclusiva das operadoras de telefonia. Contudo, através de estudos anteriores com registros de ligações telefônicas mostram enormes oportunidades como no combate a epidemias, melhoria nos serviços e na qualidade da rede de telecomunicações, planejamento de cidades inteligentes, compreensão (entendimento) do comportamento humano, entre outras importantes descobertas. Essa polêmica sobre o uso de CDR e o direito a privacidade é um debate constante, mesmo quando os pesquisadores substituem os números de telefones das pessoas por valores numéricos sequenciais, técnica chamada de anonimização. De fato, manter os dados dos usuários em sigilo é mais do que necessário e imperativo. Isso reforça a importância do objetivo deste trabalho que é apresentar uma técnica que permita realizar estudos e pesquisas com CDR preservando a restrição da privacidade das pessoas.

1.3 OBJETIVO

Para atingir esses objetivos, foi necessário investigar a seguinte pergunta: “Como trabalhar ou manipular os dados de CDR, com toda essa riqueza de informação, sem violar a privacidade das pessoas? ”

Apresentar uma técnica que permita realizar estudos e pesquisas com CDR, preservando a restrição da privacidade pessoal e propor método prático do início ao fim da implementação, visando a caminhos alternativos para planejamento de rede e/ou mobilidade urbana.

1.4 VISÃO GERAL DA OBRA

O capítulo segundo apresenta a importância do estudo de CDR, desde o surgimento, o conceito, sua importância e a interpretação dos dados e seus significados. No capítulo terceiro, é apresentada a revisão da literatura, com exemplos práticos do uso de CDR no mundo, além das limitações, como questões de privacidade das pessoas em relação ao uso de CDR. Será traçado, ainda, um quadro téorico da análise da literatura publicada sobre o tema desta dissertação e a estruturação conceitual que dará sustentação ao desenvolvimento da pesquisa. Para tanto, será abordado o que já foi publicado sobre o tema e o problema de pesquisa escolhida, permitindo um mapeamento de quem já escreveu e o que já foi escrito sobre o tema e/ou problema da pesquisa. Refere-se, também, à abordagem normativa, por meio de leis governamentais, quanto ao aspecto prático do dia a dia. O capítulo quarto descreve a fundamentação teórica – o método ***W****ork and* ***H****ome* ***E****xtracted* ***RE****gions* (WHERE) – para estudar o deslocamento de grande quantidade de pessoas dentro de diferentes áreas de uma cidade, introduzindo o principal conceito desta dissertação, que é o CDR sintético (hipotético). O capítulo quinto discorre sobre o exercício prático da metodologia e geração do CDR sintético, apresentando a metodologia utilizada nesta dissertação quanto à abordagem e implementação de CDR sintético (hipotético). Além disso, expõe os *softwares* utilizados para demonstrar a visualização do deslocamento urbano de pessoas ao longo da cidade sintética (hipotética). O capítulo sexto apresenta as considerações finais e os rumos pós-defesa.

2 REGISTRO DE LIGAÇÕES TELEFÔNICAS (CDR)

Atualmente, com o surgimento da internet e das tecnologias de Voz sobre IP (VoIP), o conceito de telefonia é ainda mais abrangente, pois envolve não somente o ato de se comunicar por meio de um telefone, mas também a troca de dados dos mais diversos (incluindo voz, vídeo, mensagens de texto, arquivos etc). Ainda, os dados das chamadas, que são gerados durante a comunicação, seja por celulares, computadores ou telefones convencionais são denominados *Call Detail Record* (CDR) (FARIA, 2010). Os CDRs são coletados e utilizados, principalmente, pelas operadoras, para que seja possível efetuar uma cobrança pelo serviço prestado ao assinante (pessoa que utiliza a rede para se comunicar); no entanto, não estão restritos à emissão de faturas; eles contêm informações da rede (como, por exemplo, motivo de falha para uma chamada não completada, rotas de comunicação utilizadas etc).

Estes registros de ligações telefônicas, CDRs, são gerados, automaticamente, nas centrais de telefonias, sejam elas de rede fixa, móvel ou IP. Esses registros são resultados de ligações telefônicas de voz ou de acesso de dados, que ocorrem em tempo real quando as ligações de voz ou de acesso de dados são realizados. Os CDRs incluem informações do horário inicial da ligação, do horário final da ligação, quantidade de dados transmitidos ao longo da comunicação, por quais centrais telefônicas a ligação de voz ou conexão de dados transitaram pelas redes de telecomunicações, informações sobre o *status* de estabelecimento ou da interrupção da conexão de voz ou de dados. Os códigos de sinalização OK ou NOK são muito relevantes porque informam o que aconteceu com a conexão de voz ou de dados. Além dessas informações, outros dados de sinalização também são registrados.

Segundo Faria (2010), a base de CDR consiste em uma das maiores fontes de informação disponibilizadas atualmente – o sistema de telefonia é, ainda hoje, o principal meio de comunicação entre as pessoas (seja para fins familiares, negócio, governamental, relacionamentos etc). Tal sistema, apesar de estar migrando, de forma acelerada, para o uso da rede de dados da internet, não vai perder sua principal característica – que consiste na conexão entre duas ou mais pessoas para a troca de informação/mídia (voz, vídeo, fotos etc). Esta transformação da telefonia para a rede de internet deve trazer ainda mais assinantes para a comunidade, o que, com certeza, vai

aumentar a geração de dados de registro de CDR dessas chamadas e permitir uma análise mais detalhada.

2.1 INTERPRETAÇÃO DA ESTRUTURA DE DADOS

Os processos de bilhetagem, de cobrança, de faturamento e de interconexão entre redes depende da interpretação dos dados contidos no CDR. A Figura 2 mostra o *layout* adaptado de um formato de CDR utilizado para bilhetagem de dados. O CDR possui diversos campos, mas o Quadro 1 apresenta apenas alguns, a título de exemplo. Outros campos tão ou mais importantes não foram apresentados porque cada fabricante escolhe um conjunto de campos para a montagem do bilhete de CDR.

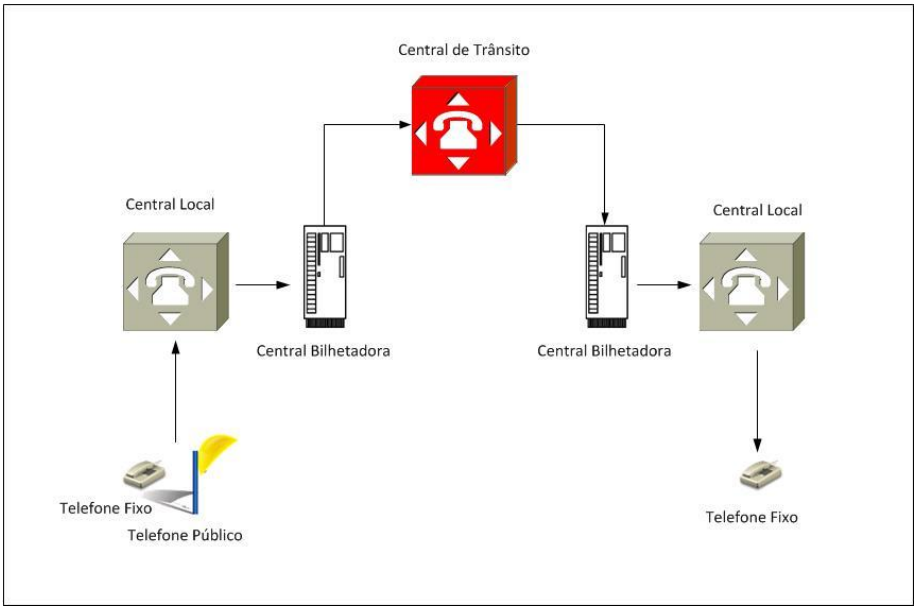


Figura 2 - Topologia básica da rede fixa
Fonte: Elaborado pelo autor[14], [15], [16]

A ligação de origem, na Figura 2, será realizada por dois tipos de terminais: (i) telefone fixo residencial[17] para telefone fixo residencial e (ii) telefone de uso público (TUP[18]) para um telefone fixo residencial. Todos os telefones envolvidos neste exemplo

[14] Central Local: É a central a que estão conectados os assinantes de uma rede telefônica em uma região.
[15] Central Bilhetadora: Central que gerou o registro da chamada (CDR)
[16] Central Trânsito: É uma central telefônica que troca informações entre outras centrais, que podem ser centrais locais ou, até mesmo, uma outra central trânsito.
[17] Classe residencial: classe de assinante de acesso individual destinado para uso estritamente doméstico.
[18] TUP: telefone de uso público – orelhões.

pertencem à mesma operadora. Os terminais têm a idêntica localização geográfica, o que estabelece e define como sendo uma chamada local[19].

O processo inicia-se na origem, com os terminais fixo residencial e TUP. Quando discado o número do destino, a central local identifica a chamada e verifica se o número de destino é atendido por ela ou não. Ao encaminhar a ligação para a próxima central, denominada de central bilhetadora[20], a central gera o registro desta chamada e, em seguida, encaminha para a central de trânsito, também conhecida como central PAS. A central de trânsito direciona a chamada para a central bilhetadora, que encaminha a chamada para a central local, que será a central do destino, cujo número de destino está conectado. Ao final do processo, a ligação é estabelecida fim-a-fim.

As informações apresentadas nos Quadros 1 e 2 são geradas por centrais telefônicas de telefonia de rede fixa.

Item	Descrição	Tipo
1	Código Único	Inteiro
2	Protocolo de sinalização	Caractere
3	Número do chamador	Caractere
4	Número chamado	Caractere
5	Identificação da Central	Inteiro
6	Endereço de origem da central	Inteiro
7	Endereço de destino da central	Inteiro
8	Identificação do canal de voz	Inteiro
9	Rota de encaminhamento da chamada	Inteiro
10	Data/Hora da chamada de voz	Data/hora
11	Duração da chamada inteira de voz	Inteiro
12	Duração da conversação	Inteiro
13	Fim de seleção	Caractere

Quadro 1 - Campos de CDR de rede de telefonia fixa
Fonte: Elaborado pelo autor

[19] Área local: área geográfica contínua de prestação de serviços, definida pela ANATEL, segundo critérios técnicos e econômicos, onde é prestado o Sistema Telefônico Fixo Comutado (STFC) na modalidade local.

[20] Central que gera chamada bilhetada. Tipo de chamada cujos atributos – código de acesso e categoria do assinante chamador, código de acesso e sinal de fim de seleção do assinante chamado, data, hora de início, duração – são registrados de forma individualizada.

Identificação do campo	Descrição
Código Único	Este campo é único e diferencia cada linha de registro de CDR de outro.
Protocolo de sinalização	Informa o protocolo que está sendo utilizado pela central telefônica de rede fixa.
Número do chamador	Identifica o terminal de origem da chamada de voz.
Número chamado	Identifica o terminal de destino da chamada de voz.
Identificação da Central	Este campo define a central bilhetadora que gerou os registros de CDR.
Endereço de origem da central	Identificação de origem da central telefônica para a comunicação de voz entre centrais.
Endereço de destino da central	Identificação de destino da central telefônica para a comunicação de voz entre centrais.
Identificação do canal de voz	Especifica o canal lógico utilizado para que a voz seja encaminhada.
Rota de encaminhamento da chamada	Identifica o caminho que a chamada de voz percorre ao longo do trajeto de comunicação.
Data/Hora (HH:MM:SS) da chamada de voz	Horário em que a comunicação se iniciou.
Duração da chamada inteira de voz	Tempo total da chamada de voz, que inclui informações adicionais de sinalização da conexão, em segundos.
Duração da conversação	Tempo total somente da comunicação de voz, em segundos.
Fim de seleção	Identifica a causa da conexão de voz, registrando informações referentes aos terminais de destinos e contém informações de chamadas OK e NOK (chamadas que apresentaram falhas).

Quadro 2 - Descrição resumida dos campos exibidos no Quadro 1
Fonte: Elaborado pelo autor

A Figura 3 representa uma topologia genérica da rede móvel de telefone.

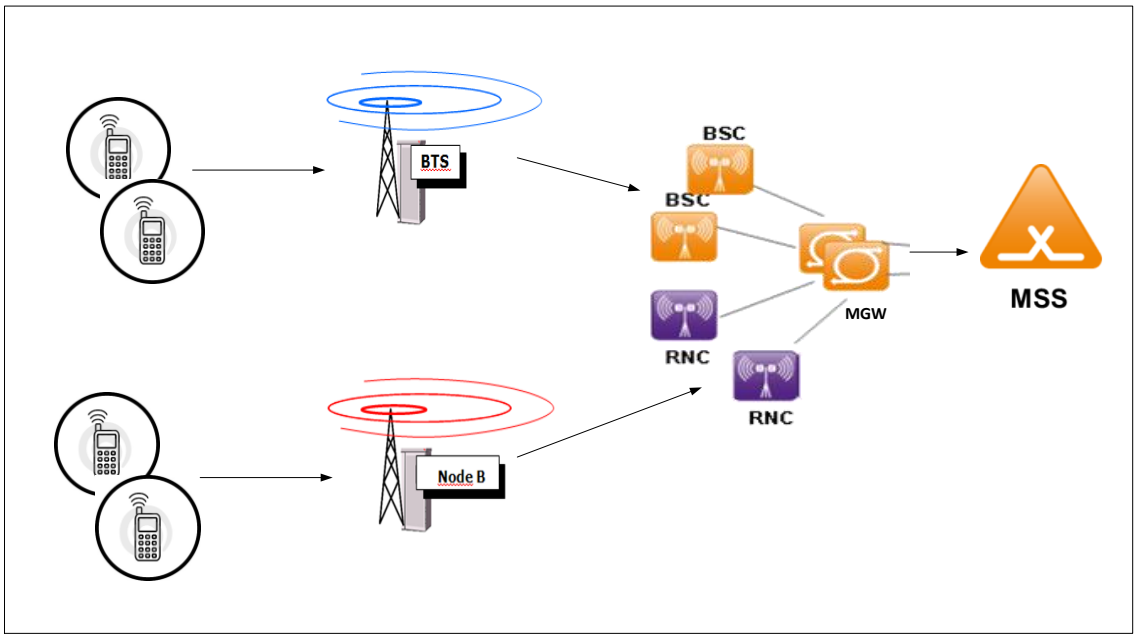


Figura 3 - Topologia básica da rede móvel
Fonte: Elaborado pelo autor[21], [22], [23], [24], [25]

O telefone celular estabelece a comunicação entre os usuários e a rede móvel. A primeira troca de mensagem, dependendo da tecnologia de acesso, pode ser com a Base Transceiver Station (BTS). As BTSs são controladas pelas Base Station Controller (BSC). Uma BSC controla várias BTSs. As principais funções da BSC são as de controlar, gerenciar e monitorar as BTS.

A rede de acesso do 3G é chamada de Universal Terrestrial Radio Access Network (UTRAN), que é composta da controladora da rede rádio – Radio Network Controller (RNC) – e da Node B. A RNC controla uma ou mais Nodes B. A Node B tem finalidade semelhante a uma BTS. Tanto a BSC quanto as RNCs estão conectadas ao elemento de rede chamado Media Gateway (MGW). As Medias Gateways estão

[21] BTS: Base Transceiver Station, de acordo com o documento Technical Specification: 3GPP TS 23.002 V13.1.0 (2014-12), p. 34.
[22] BSC: Base Station Controller, de acordo com o documento Technical Specification: 3GPP TS 23.002 V13.1.0 (2014-12), p. 20.
[23] RNC: Radio Network Controller, de acordo com o documento Technical Specification: 3GPP TS 23.002 V13.1.0 (2014-12), p. 20.
[24] MGW: Media Gateway, de acordo com o documento Technical Specification: 3GPP TS 23.002 V13.1.0 (2014-12), p. 27.
[25] MSS ou MSC: Mobile switching center server, de acordo com o documento Technical Specification: 3GPP TS 23.002 V13.1.0 (2014-12), p. 26.

conectadas ao núcleo da rede, no elemento chamado de Circuit Switched (CS - comutação por circuito), para as chamadas de voz e videochamadas, e no domínio de comutação por pacote Packet Switched (PS) para chamadas de dados. A central de comutação na rede núcleo é mais conhecida por Mobile Switching Center (MSC).

As informações apresentadas no Quadro 3 são geradas por centrais telefônicas de telefonia móvel.

Número sequencial	Nome do campo	Descrição
1	Número de A	Número do chamador
2	IMEI do número do chamador	Identificação Internacional de Equipamento Móvel. É um número único que identifica cada aparelho de telefone celular.
3	IMSI do número do chamador	É o número que identifica o usuário à operadora de telefonia móvel.
4	Número de B	Número do telefone chamado.
5	IMEI do número do número chamado	Identificação Internacional de Equipamento Móvel. É um número único que identifica cada aparelho de telefone celular.
6	IMSI do número chamado	É o número que identifica o usuário à operadora de telefonia móvel.
7	Data e Hora iniciada	Chamada iniciada
8	Data e Hora da terminada	Chamada terminada
9	Cell ID	Setor em que a chamada iniciou.
10	Duração	Duração da chamada (em segundos)
11	LAC Cell ID inicial	Este código é utilizado para a localização da área inicial do assinante chamador.
12	Cell ID	Setor em que a chamada terminou.
13	LAC Cell ID final	Este código é utilizado para a localização da área final do assinante chamado.
14	First Cell ID inicial	Setor em que a ligação iniciou.
15	Cell ID final	Setor em que a ligação terminou.
16	Clear Code	Identifica a causa de interrupção da chamada.
17	Tecnologia	Identifica a tecnologia da central que originou a chamada.

Quadro 2 - Campos de CDR de rede de telefonia móvel
Fonte: Elaborado pelo autor

2.2 APLICAÇÕES DE CDR EM TELECOMUNICAÇÕES

Segundo o Portal de informações do setor de telecomunicações do Brasil (TELECO, 2015), os CDRs são utilizados pelo sistema de bilhetagem – processo no sistema telefônico que permite a aquisição e gravação de informação sobre as chamadas tais como: quem originou/recebeu, onde, quando, por quanto tempo etc. – e pelo sistema de tarifação – sistema que se propõe a associar um custo correspondente a todas as ligações recebidas (entrantes) e efetuadas (saíntes).

Nesse contexto, os sistemas pertencentes ao ciclo de receita da operadora são considerados como ponto central para o suporte ao negócio da companhia (COSTA, 2010). Esses sistemas são responsáveis pelo processamento de todos os serviços

solicitados pelos clientes, desde o uso de serviços individuais até a geração da fatura telefônica, contemplando todos os planos de preços e tarifas comercializados com cada um de seus clientes. Segundo o autor, pesquisas divulgadas mundialmente mostram que as operadoras perdem de 1 a 3% de suas receitas devido a dificuldades operacionais, que vão desde problemas nos elementos de rede (centrais e plataformas) até os sistemas de informação (Mediação[26], *Rating*[27], *Billing*[28], Interconexão[29], *Cobilling*[30], Arrecadação[31] e Cobrança[32]). Todos esses sistemas utilizam a base de CDR, que tem todas as informações dos clientes[33].

Outra oportunidade a ser explorada consiste na análise de CDRs gerados em sistema de *contact center* (onde são realizadas vendas, *marketing* e atendimento de reclamação dos clientes). É possível utilizar os CDRs destes sistemas para que sejam encontrados padrões de comportamento tanto dos clientes que se relacionam com a empresa quanto dos funcionários responsáveis em atendê-los com qualidade (FARIA, 2010).

O monitoramento do perfil de tráfego de uso do serviço e do cliente para enviar notificações ou alertar quando o assinante está diminuindo em muito o uso da rede (o que pode significar uma mudança de operadora ou redução do custo por algum motivo) é outra aplicação oriunda da análise de CDR. Do mesmo modo, é possível alertar quando um assinante passa a utilizar a rede em demasia, extrapolando seu comportamento natural ou seu consumo habitual (o que pode significar fraude ou clonagem do aparelho).

O CDR também vem sendo utilizado para diversas aplicações nas operadoras de telecomunicações. Em função do alto volume de dados que são gerados instantaneamente pelos CDRs (registros das ligações telefônicas e de acessos de dados), as operadoras utilizam essa enorme base de dados para: (i) avaliar o desempenho da rede de telecomunicações (interconexão entre centrais telefônicas); (ii) auditoria da rede

[26] Mediação – é o processo de análise dos CDR, gerados na rede, visando sua classificação em descartados, de faturamento de público, de faturamento de interconexão, etc...
[27] Rating – é o processo de tarifação, ou seja, valorar o serviço do respectivo CDR.
[28] Billing – é o processo de faturar o conjunto de serviços utilizados pelo cliente ao longo do seu ciclo de faturamento.
[29] Interconexão – é o processo de seleção dos CDR que pertencem às rotas de interconexão visando gerar o encontro de contas entre as operadoras envolvidas.
[30] Cobilling – é o processo entre operadoras para inserir na fatura de seus clientes os serviços que eles utilizaram de outras operadoras.
[31] Arrecadação – é o processo de controle do recebimento das faturas emitidas.
[32] Cobrança – é o processo de atuação junto aos clientes inadimplentes.
[33] Informações dos clientes são telefone de origem e telefone de destino.

de telefonia, identificando falhas de sinalização[34] entre as principais centrais de telefonia ou elementos de rede; (iii) auditoria financeira, provendo meios para análise de perda/evasão de receitas; (iv) congestionamento de tráfego; (v) falhas na formação do bilhete de CDR e (vi) identificação de chamadas fraudulentas. A monitoração de todos esses serviços permite um melhor gerenciamento da qualidade da oferta dos serviços de telecomunicações e da qualidade da rede de telecom (QoS). As informações da base de dados de CDR gera conteúdo de alta relevância para diferentes áreas da operadora.

Yan, Fassino e Baldasare (2005) asseveram que a inteligência artificial tem sido utilizada para prever o comportamento dos assinantes, na indústria de telecomunicações. O CDR, em particular, por ser uma fonte de dados rica em informação do faturamento, serviços utilizados pelos assinantes, informações geográficas, pode ser utilizado, também, em um modelo preditivo de *churn*[35] (índice de cancelamento de clientes), que é a mudança dos assinantes de uma operadora para outra operadora – ou melhor, saída do assinante de uma base de clientes de uma operadora para outra operadora. O mercado de telecomunicações está tão competitivo que as operadoras buscam realizar pesquisas sobre o comportamento da sua base de assinantes, com o objetivo de predizer o que vai acontecer no futuro, principalmente quando se trata de perdas de assinantes. Segundo estudos realizados por Michael et al. (2000), um dos fatores que mais influenciam os assinantes a trocarem de prestadoras de serviços de telecomunicação é a qualidade do serviço de voz e de dados. Esse fator tem importância da ordem de 21% para usuários realizarem *churn*. Os altos índices de insatisfação registrados nos órgãos de defesa do consumidor são alarmantes.

Por isso, o *churn* é um índice muito importante a ser acompanhado pelas áreas de negócios dessas empresas, bastante conhecido, e serve para realização de estratégias ou campanhas de *marketing* na retenção da base de clientes.

O estudo realizado por Yan, Fassino e Baldasare (2005) revela que, nos Estados Unidos e na Europa, o *churn* representa um valor significativo de perda financeira da ordem de US$4 bilhões por ano. Neste sentido, o investimento na pesquisa e na predição do comportamento do assinante é uma premissa estratégica, a fim de que essas empresas entendam as razões que provocam o *churn*, permitindo monetizar e reter a base de clientes.

[34] Por sinalização SS7 entende-se o conjunto de mensagens trocadas sequencialmente entre centrais telefônicas e responsáveis pelo estabelecimento de uma chamada.
[35] Churn: termo que designa a rotatividade dos usuários dos serviços de uma empresa para outra.

A inovação desse estudo foi tratar a base de clientes pré-pagos e utilizar dados de CDR como fonte de dados. Através da base de CDR, foi possível identificar o número para o qual a pessoa mais liga, a frequência com que essas pessoas realizam as ligações para as mesmas pessoas, quem liga para quem e fazer cruzamentos desses dados. Por um período de seis meses de acompanhamento da base de CDR, foi possível identificar as principais causas das mudanças de operadoras.

Outro estudo de caso relevante sobre *churn*, realizado pela operadora brasileira de telecomunicações Oi foi apresentado por um de seus executivos, Raphael Stein, na conferência Partners User Group 2012, em Washington. Tadeu (2012) relata que a Oi desenvolveu um projeto de análise de redes sociais baseado nas relações de tráfego dentro da sua própria rede, com o propósito de reduzir a taxa de desconexão (*churn*), migrando para outra operadora.

A ideia do projeto foi desenvolver um modelo baseado em métricas para avaliar o impacto do *churn* viral. Foi possível descobrir o líder da rede social, ou seja, o cliente que possui capacidade de influenciar outras pessoas ou grupos que usam os serviços da operadora. O perfil do líder foi definido com base no grau de influência sobre os demais usuários da rede, a frequência de uso dos serviços, o tempo de conversação e o Average Revenue Per User (ARPU)[36] elevado, entre outros aspectos.

A Oi desenvolveu o roteiro de métricas que avaliam, por exemplo, o número de incidentes com conexões, a porcentagem de conexões à rede e sua relevância, o tempo gasto por outros nós da rede ao nó do líder, frequência, duração, o controle das chamadas em vários momentos do dia, se a ligação é local, nacional ou internacional, entre outros itens. Com isso, é possível saber quais são os clientes que influenciaram outros a deixarem a operadora.

A técnica de deslocamento das pessoas, utilizada na produção de CDR sintético apresentada neste trabalho, prevê o sorteio da população e de residências e postos de trabalho. Em seguida, defina-se a classe de ligadores, que estabelece a frequência que cada tipo de pessoa realiza em suas ligações telefônicas diariamente. Utilizar a probabilidade para calcular a quantidade de ligações diárias realizadas pelas pessoas e, a partir desse cálculo, identificar de onde essas ligações foram realizadas podem simular um cenário real do dia a dia. A localização estimada da residência e do trabalho ocorre por meio das chamadas realizadas no mesmo horário diariamente, o que estabelece um

[36] Average Revenue Per User (ARPU) significa a receita média por cliente (ou usuário).

padrão. Já a identificação geográfica do local presencial em que a pessoa se encontra é fornecida pela localização da torre de celular, a qual a chamada foi realizada e/ou recebida.

Considera-se no estudo de mobilidade dois principais fatores: (i) o tempo (horário que as pessoas se movimentam) e (ii) o espaço (informações sobre lugares geográficos que os indivíduos mais frequentam em seu cotidiano, especialmente, a residência e o trabalho). Nesse sentido, entender onde as pessoas passam o seu tempo é importante para o estudo e proposição de novas formas de mobilidade em grandes centros urbanos. Segundo estudos realizados por Isaacman (2012), a maior parte das pessoas, prioritariamente, gasta seu tempo tanto em suas residências quanto em seus trabalhos. Por utilizar o aparelho de celular para realizar e receber ligações nesses lugares, os indivíduos podem ser facilmente localizados fisicamente, principalmente ao longo de um horário definido de trabalho das 9 às 18 horas.

É importante ressaltar que nem todos os pesquisadores têm acesso aos CDRs reais, porque os CDRs não são facilmente disponíveis. Por este motivo, a alternativa do método de produção de CDR sintético utiliza os dados populacionais de censo geográfico para reproduzir, similarmente, as mesmas condições de estudos reais.

O maior benefício do uso de CDR sintético é ter a possibilidade de predizer e visualizar o impacto das mudanças hipotéticas no estudo de mobilidade regional. O estudo desses impactos, por exemplo, contribui para a melhor distribuição e reordenação de novos lugares residenciais e criação de postos de trabalhos em novos locais, estabelecendo melhor harmonia com a mobilidade urbana, criada a partir de uma simulação em ambiente “sintético” (hipotético). Outro importante benefício do uso de CDR sintético é relacionado com a preservação da privacidade das pessoas, quando se estuda o deslocamento urbano dos indivíduos. Enquanto no uso do CDR sintético o número do telefone é sintético (artificial), no estudo com CDRs reais os números são também reais (verdadeiros), colocando em risco a revelação da identidade dos indivíduos e seus respectivos hábitos cotidianos. Outra vantagem no uso do CDR sintético é utilizar os mesmos *softwares* (programas de computadores) que são utilizados pelos CDRs reais. Esta é mais uma forma de se comparar os resultados e verificar que são equivalentes.

3 REVISÃO DA LITERATURA

Diversos estudos sobre o uso de registros de ligações telefônicas vêm sendo realizados por pesquisadores no mundo. Este capítulo traz uma versão da literatura sobre este assunto.

As riquezas de dados contidos nos CDRs móveis permitem acompanhar a trajetória das pessoas ao longo do dia. Associar padrões de comportamento a estas trajetórias permite entender melhor a mobilidade humana. É importante ressaltar que os telefones móveis estão se tornando cada vez mais onipresentes em toda a parte do mundo, especialmente em áreas urbanas densamente povoadas e, em particular, nos países industrializados, onde a penetração de telefonia móvel é de quase 100%. Provedores de telefonia móvel coletam, diariamente, grande quantidade de dados das chamadas telefônicas. Esses registros de CDR podem revelar informações importantes para contribuir no planejamento urbano. Em geral, esses estudos contribuem no entendimento de epidemias e no planejamento da mobilidade urbana.

3.1 PADRÕES DE MOBILIDADE DAS PESSOAS A PARTIR DO USO DE CDRs

Identificar padrões de mobilidade das pessoas é fundamental para a compreensão mais profunda dos efeitos do comportamento humano. Poder utilizar os dados de CDR para identificar os padrões de comportamento das pessoas é um recurso tecnológico essencial nos dias de hoje. Song et al. (2010) demostraram, em seu estudo, que apesar de diferentes indivíduos apresentarem comportamentos bem distintos, essas pessoas tendem a gastar mais de seu tempo nos lugares onde mais frequentam. A entropia[37] é, provavelmente, o fundamento mais importante para inferir sobre o destino ao qual a pessoa poderá estar. No estudo, Song et al. (2010) atribuíram três tipos de entropia ao padrão de mobilidade das pessoas para estudar seu comportamento nos deslocamentos. A primeira foi a entropia aleatória, que foi inserida com o objetivo de capturar o grau de previsibilidade de identificar o possível destino dos indivíduos. A segunda foi a entropia temporal, que correlaciona a probabilidade histórica (ou padrão

[37] A entropia representa a probabilidade de se identificar uma sequência ou a trajetória natural de uma pessoa qualquer, o que pode ser mensurado através do grau de irreversibilidade de um sistema ou da propriedade de um sistema sofrer alterações.

histórico) de localização do indivíduo com o padrão de visitação usual das pessoas (padrão pré-identificado). A terceira foi a entropia real, que depende da frequência de visitação do local (do destino), mas que também considera o tempo gasto pelas pessoas nesse lugar de destino (local visitado). Essas perturbações construíram um espaço temporal completo e um padrão de mobilidade humana. O estudo constatou que os indivíduos que "viajam" menos, ou melhor, que se deslocam menos, são mais fáceis/previsíveis em seus deslocamentos de destinos, tendo em vista a pequena entropia inserida nesse cenário. Em compensação, aqueles com grandes deslocamentos devem ser mais difíceis de prever seus destinos, em função da alta entropia inserida neste cenário. Outra importante conclusão do trabalho realizado por Song et al. (2010) foi a constatação da possibilidade de prever o destino de um usuário.

Para Becker et al. (2011), compreender o padrão de mobilidade das pessoas que vivem e utilizam a cidade é de alta relevância. O melhor entendimento da mobilidade das pessoas facilita criar soluções para problemas de congestionamento de tráfego, vagas de estacionamento de veículos em grandes centros urbanos, seguranças das pessoas e outros importantes aspectos da vida urbana. Em seu estudo sobre padrões de *handoff*[38] para celulares, Becker et al. (2011) identificaram as rotas (ruas, avenidas e rodovias) que as pessoas trafegaram dentro da cidade, ao longo de seus movimentos em ônibus, em carros, em trens e em outros meios de transporte. Para Isaacman et al. (2010), redes de telefonia celular podem ajudar a resolver problemas importantes fora do domínio das comunicações, porque elas fornecem informações valiosas sobre a forma de como as pessoas se deslocam ao longo do dia. Além disso, os cientistas e formuladores de políticas públicas, em muitos campos, podem usar a mobilidade humana para explorar os problemas existentes e antecipar futuros problemas. Isaacman et al. (2010), em seu estudo, procuraram entender as diferenças de comportamento das pessoas de duas grandes populações: Los Angeles e Nova Iorque. Ao analisar os registros anônimos de celular (CDR), foi possível extrair conclusões sobre como as pessoas se movem em torno de duas grandes cidades dos Estados Unidos, em particular.

O estudo realizado por Mafla Sánchez (2013) sobre mobilidade e transporte público na cidade contemporânea apresenta diversos significados para o entendimento sobre mobilidade e acessibilidade. Uma das definições apresentada consiste na

[38] *Handoff* é a sequência de antenas celulares que, ao longo de uma chamada telefônica de voz, o telefone é registrado (ou "acampado").

mobilidade ser um dos valores mais importantes da sociedade contemporânea, porque trata de elemento fundamental da dinâmica demográfica, já que congrega uma série de fenômenos imprescindíveis para compreender as transformações no mundo atual. Ainda segundo Sánchez (2013), o conceito de mobilidade está relacionado com os deslocamentos diários (viagens) de pessoas no espaço urbano. Não apenas a sua efetiva ocorrência, mas também a facilidade e a possibilidade de ocorrência. Neste sentido, a agilidade em coletar informações mais rapidamente sobre o deslocamento de pessoas é fundamental. O trabalho realizado por Hanson et al. (2011) reflete essa cogitação, pois traz uma importante análise comparativa sobre mobilidade urbana utilizando registros de ligações telefônicas (CDRs) *versus* informações públicas de censo de pesquisa dos Estados Unidos – *2000 Census Transportation Planning Package (CTPP)* –, que inclui informações específicas sobre o deslocamento das pessoas entre o trabalho e suas residências. Para Hanson et al. (2011), os estudos tradicionais sobre mobilidade urbana são caros e envolvem técnicas de contagem de veículos. Além disso, estudos de mobilidade pendular de grande dimensão (grande escala) demoram anos para serem concluídos. Em muitos casos, o planejamento de muitos municípios somente é iniciado após a publicação das pesquisas pelos censos. Em contrapartida, utilizar oportunisticamente a rede de telefonia celular, através de CDR, para identificar locais de grandes movimentos de pessoas, é um diferencial extraordinário na dinâmica do planejamento dos centros urbanos. Deve-se considerar que os telefones celulares são onipresentes e, por sua proximidade quase que constante com seus usuários, acabam se tornando um sensor muito eficiente. Além disso, o uso do CDR para estudar a dinâmica das cidades é infinitamente mais barato, mais rápido e tão confiável como as pesquisas realizadas pelos institutos públicos de pesquisas de geografia e estatística.

Para Silva (2014), o deslocamento "casa⇔trabalho⇔casa" fornece informações relevantes sobre o padrão de viagem “casa⇔trabalho⇔casa”. A maioria dos deslocamentos com propósito "trabalho" é realizada, rotineiramente, pelas mesmas pessoas, para os mesmos locais e, em geral, nos mesmos horários em dias úteis. A partir dessa conclusão, é possível a obtenção de dados suficientes para a identificação de pontos críticos e de proposição de alternativas que desestimulem o uso do transporte individual motorizado. Segundo esse autor, um dos principais sintomas da piora das condições de mobilidade refere-se à quantidade de tempo que a população "desperdiça" no trânsito. Outro estudo complementar, realizado por Teerayut (2012) sobre o percurso das pessoas da “casa⇔trabalho⇔casa” analisa o deslocamento das pessoas em seu

cotidiano, considerando os lugares mais frequentados ao longo do trajeto diário. A Figura 4 caracteriza o percurso desse trajeto.

Figura 4 - Trajeto de uma pessoa em um dia
Fonte: Teerayut (2012)

Legenda:
TRIP 1: horário inicial do deslocamento
TRIP 2: origem do deslocamento
TRIP 3: meios de transportes (ex. trem, ônibus, carro)
TRIP 4: motivo do deslocamento (ex. trabalho, shopping, estudo)
TRIP 5: destino do deslocamento
TRIP 6: horário final do deslocamento

A pesquisa realizada por Teerayut (2012) é considerada como modelo de referência, porque compara os resultados de registros de chamadas telefônicas móveis (CDR) *versus* pesquisa obtida *online* de questionário sobre o uso do telefone celular durante o deslocamento da pessoa ao longo de seu cotidiano. O formulário foi respondido por 1.063 pessoas, enquanto o volume de CDR coletado foi de 1.260.000.000 de registros.

O Quadro 4 apresenta um resumo das principais características e hábitos de uso de celular das pessoas que responderam ao questionário da pesquisa.

Comportamento	⇨ 62% das pessoas nunca desligam seus celulares
	⇨ 10% das pessoas desligam seus celulares quando dormem
	⇨ 7% das pessoas desligam seus celulares quando estão trabalhando
	⇨ 62% das pessoas mantêm seus celulares próximos
	⇨ 30% das pessoas mantêm seus telefones ao lado delas 24 horas por dia
	⇨ 60% das pessoas tendem a utilizar o telefone celular quando estão em engarrafamentos
	⇨ 5% das pessoas desligam seus celulares quando estão em deslocamento
	⇨ 20% das pessoas desligam seus celulares em lugares específicos como teatro, cinemas e *shows*
	⇨ 37% das pessoas utilizam seus telefones celulares para o serviço de mensagem de texto
	⇨ 18% das pessoas utilizam o serviço de internet
	⇨ 13% das pessoas utilizam seus telefones celulares para realizarem ligações telefônicas
Frequência de ligações	⇨ 33% das pessoas realizam poucas ligações telefônicas por semana
	⇨ 26% das pessoas realizam uma ou duas ligações telefônicas por dia
	⇨ 25% das pessoas realizam algumas ligações telefônicas por mês
	⇨ 12% das pessoas realizam de três a seis ligações por dia
	⇨ 4% das pessoas realizam até seis ligações por dia
Duração da ligação	⇨ 63% das pessoas falam ao telefone móvel em média de 1 a 5 minutos
	⇨ 19% das pessoas falam ao telefone móvel de 5 a 15 minutos
	⇨ 10% das pessoas falam ao telefone móvel menos que 1 minuto
Uso do celular no trabalho	⇨ 21% das pessoas utilizam os telefones no trabalho
	⇨ 19% das pessoas não utilizam telefone no trabalho
	⇨ 15% das pessoas utilizam o telefone no trabalho algumas vezes por semana
	⇨ 13% das pessoas utilizam o telefone no trabalho 1 ou 2 vezes por dia
	⇨ 6% das pessoas utilizam o telefone no trabalho de 3 a 6 vezes ao dia
Uso do celular no deslocamento para o trabalho	⇨ 31% das pessoas nunca falam ao telefone móvel
	⇨ 19% das pessoas falam ao telefone móvel quando estão em deslocamento
	⇨ 13% das pessoas falam ao telefone móvel pelo menos uma vez por semana
	⇨ 9% das pessoas falam ao telefone móvel pelo menos 1 a 2 vezes por dia
	⇨ 25% das pessoas não precisam se deslocar para o trabalho diariamente (exemplo: trabalho em casa)
Uso do celular na residência	⇨ 35% das pessoas fazem ligações telefônicas algumas vezes por mês
	⇨ 31% das pessoas fazem ligações telefônicas algumas vezes por semana
	⇨ 21% das pessoas fazem ligações telefônicas uma ou duas vezes ao dia
	⇨ 6% das pessoas fazem ligações telefônicas três a seis vezes por dia
	⇨ 5% das pessoas não fazem ligações telefônicas em casa

Quadro 4: Perfil de uso do celular

Fonte: Teerayut (2012)

A contribuição do estudo realizado por Teerayut (2012) foi importante para confirmar o potencial uso dos registros de CDR como importante insumo para compreender o deslocamento das pessoas e o uso do celular como sensor de localização. Marques Neto et al. (2013) publicaram um trabalho sobre a mobilidade humana em eventos de larga escala, utilizando registros de chamadas telefônicas (CDRs). Esse estudo foi um dos primeiros trabalhos realizados no Brasil dedicado a analisar e a entender o deslocamento das pessoas em eventos de grande público, como jogos do campeonato Brasileiro de Futebol de 2011 e durante algumas comemorações de *réveillon* de 2011-2012, a partir de registro de chamadas telefônicas de voz realizadas em uma rede de telefonia celular. Para Marques Neto et al. (2013), os padrões de comportamento de mobilidade humana, inferidos a partir do histórico de uso de serviços de telefonia móvel, podem apoiar o desenvolvimento de melhores estratégias de gestão de recursos e de serviços de comunicação, tanto para as pessoas que frequentam esses eventos de grande porte, quanto para aqueles que vivem nessas cidades. Alguns questionamentos que foram endereçados nesse trabalho – e que estão diretamente associados à mobilidade urbana – são: Quem esteve nos arredores dos locais da realização do evento? De onde essas pessoas vieram? E para onde eles se deslocaram após o evento?

Mais recentemente, em entrevista a Francisco Medeiros, Pablo Cerdeira, coordenador do *Big Data* e cientista de dados da Prefeitura do Rio de Janeiro, descreveu a experiência com CDR realizada no *Réveillon* de 2013/2014 em Copacabana pela Prefeitura Municipal. Este trabalho foi inovador e mostrou, eficientemente, a relevância do uso de registros para cenários próximos do tempo real. O problema a ser resolvido nesse estudo foi entender como, aproximadamente, dois milhões de pessoas que passam o *réveillon* na praia de Copacabana, vindos de todas as regiões do Rio de Janeiro, conseguem retornar para as suas residências ao longo da madrugada. Para que bairros essas pessoas retornarão? Em quanto tempo? Como? Foram selecionadas todas as ligações que tiveram como destino o bairro de Copacabana ao longo do pré-evento, as ligações originadas de Copacabana durante o evento e as ligações que ocorreram com origem em Copacabana após o *Réveillon*. Desta forma, foi possível identificar de que bairros de origem são as pessoas e os respectivos bairros de destino. A relevância desse estudo foi colaborar, rapidamente, com o plano de transporte urbano que a cidade do Rio de Janeiro possui para ajudar no planejamento de escoamento dessa população. Isso

tudo foi possível porque os dados de telefonia móvel de celular ocorrem em tempo real[39] (são dinâmicos).

Para Amy Wesolowski et al. (2014), o uso de CDRs contribuiu para a contenção do surto de Ebola. Nesse estudo, os autores relataram que a rápida propagação do vírus de Guiné, Serra Leoa e Libéria para Nigéria e Senegal foi influenciada pelo deslocamento local e regional das pessoas. Por este motivo, ter a informação sobre a distribuição da população e do seu deslocamento pelas cidades é estratégico para traçar prováveis rotas de indivíduos infectados que se misturam com a população, provocando novos surtos e aumentando a transmissão do vírus. Além de entender essa movimentação da população, viabiliza-se o envio de mais profissionais a lugares mais necessitados, e pode-se melhorar a comunicação com pessoas das regiões afetadas. A riqueza dos dados contidos na base de CDR, nesse caso, permitiu que autoridades pudessem adotar medidas mais eficazes na política de vigilância de fronteiras.

Quando se trata da dinâmica humana, existe uma lacuna de conhecimento ou da incerteza entre o que é, de fato, intuição e o que é padrão de modelagem do comportamento humano. Para Song et al. (2010), o padrão de atividade do indivíduo é aleatório e imprevisível, mas fundamentalmente estocástico[40]. O uso da fórmula de *Erlang* e do modelo matemático *Lévy Walk*, utilizado em telefonia para descrever mobilidade das pessoas, são exemplos de um comportamento estocástico. Os fundamentos matemáticos ajudam a conduzir ao seguinte questionamento: Qual é o papel do acaso no comportamento humano e em que grau o indivíduo é previsível? Este estudo realizado por Song et al. (2010) teve como objetivo quantificar a interação entre o normal portanto, previsível, e o aleatório, logo, o imprevisível, utilizando, para isso, a mobilidade dos indivíduos e os limites das pessoas que caracterizam a previsibilidade da dinâmica humana. Portanto, para prever o deslocamento urbano das pessoas é necessário responder ao questionamento: até que ponto o comportamento humano é previsível? No trabalho realizado por Song et al. (2010), algumas afirmações foram surpreendentes, como, por exemplo, os fatores idade e gênero influenciavam na regularidade e previsibilidade do comportamento dos usuários. Outro fator relevante foi se a densidade populacional rural e urbana influenciaria no comportamento das pessoas,

[39] Os registros de ligações telefônicas são geralmente registrados na rede de acesso (antenas) das redes celulares. Neste caso, foram delimitadas as antenas de acessos que cobriam o evento de Réveillon de Copacabana.

[40] Qualquer tipo de evolução temporal (determinística ou essencialmente probabilística), que seja analisável em termos de probabilidade, pode ser chamada de processo estocástico.

mas também não foram identificadas variações significativas; por fim, outro fator de relevância analisado foi não terem sido identificadas mudanças significativas no comportamento dos usuários nos finais de semana quando comparado ao seu cotidiano semanal. O estudo concluiu, ainda, que a regularidade humana não é imposta pelos seus horários de trabalho, mas potencialmente está associada às atividades humanas.

Para Candia et al. (2008), o padrão de mobilidade das pessoas diminui durante a noite e, ao meio-dia, possui picos de ligações telefônicas. Isso indica que, embora as atividades das pessoas variem enormemente (ou seja, são heterogêneas), o percentual de pessoas que realizaram chamadas durante seus deslocamentos permanece estável. Essa análise contrapõe ao que Song et al. (2010) apresentaram como resultado em seu estudo. Candia et al. (2008) constataram que é possível identificar as características de padrões de perfil de ligador através do comportamento das pessoas ao longo do dia, o que é bastante razoável, tendo em vista a dinâmica do cotidiano das pessoas. Os resultados coletados confirmaram que o CDR tem um conteúdo muito rico de informação e que é uma ferramenta poderosa para entender o comportamento dos hábitos das pessoas. Toda essa informação oferece uma oportunidade para melhorar a compreensão das redes sociais e correlacionar as características comportamentais das pessoas no espaço e no tempo, mediante a distribuição de perfil de ligação telefônica. Para Wesolowski et al. (2014), o uso do CDR disponibiliza a coleta de dados dinâmicos e atualizados mais próxima do tempo real, sendo um dos maiores benefícios para uma rápida ação de combate à epidemia. Ainda segundo os referidos autores, as fontes de informação – como censo de pesquisa domiciliar – que se têm hoje em dia são limitadas e, muitas vezes, desatualizadas, o que acaba dificultando o combate à doença. Quando um novo conjunto de casos de Ebola aparece fora das áreas em atendimento, os atores de saúde não conseguem rastrear rapidamente os padrões de dispersão porque os agentes de saúde não têm quaisquer informações que mostrem, dinamicamente, a mobilidade da população ou interações sociais. Por este motivo, a importância do uso do registro de ligações telefônica (CDR) é altamente relevante.

Por este motivo, Candia et al. (2008) recomendam investigar a dinâmica de ligações dos indivíduos por dia. Estudos anteriores já haviam medido, na linha do tempo, o comportamento individual de ligações das pessoas. A frequência de ligação das pessoas pode ser estudada através do modelo de *Poisson*. Como muitas outras atividades humanas, o perfil de chamada é altamente heterogêneo também. Enquanto alguns usuários raramente utilizam o telefone celular, outros fazem centenas ou

milhares de chamadas a cada mês. Para Candia et al. (2008), diferentes técnicas podem ser utilizadas para entender a distribuição de frequência temporal das ligações telefônicas realizadas pelos usuários de telefonia móvel. A primeira técnica utilizada é (i) a distribuição *Poisson*[41]; a segunda técnica é (ii) a função de densidade de probabilidade[42] e a terceira técnica é (iii) o fundamento da física de lei de potência[43].

Para González, Hidalgo, e Barabási (2008), as pessoas seguem um padrão reproduzível no deslocamento urbano diário, em uma distribuição de probabilidade temporal. Um exemplo deste padrão pode ser observado no deslocamento das pessoas da residência para o trabalho e do trabalho para a residência. Esta sequência segue um fluxo-padrão de comportamento e pode ser compreendido utilizando o estudo de movimento aleatório de *Lévy Flight*[44]. Assim, o padrão de movimento humano pode ser explicado pelos seguintes fatores: (i) no primeiro, cada indivíduo segue uma trajetória *Lévy Flight* com distribuição de tamanho de salto independente; (ii) no segundo, a distribuição observada pode pertencer a uma base populacional heterogênica, mas que possui padrões específicos idênticos em seus comportamentos; e a (iii) terceira, pode se considerar que conjuntos de pessoas com padrões diferentes possam coexistir utilizando trajetórias individuais independentes como pode ser também explicado utilizando o método *Lévy Flight*.

No estudo realizado por Song et al. (2010), em média, 70% das vezes, o local mais visitado pelos usuários coincide com a localização real do usuário[45]. Essa afirmação pode ser comprovada durante o período noturno, no qual as pessoas passam, a maior parte do seu tempo, estacionadas. No período em torno das 13 horas, percebeu-se ser o horário usualmente de almoço (com baixo deslocamento das pessoas). Já o período do final do dia, das 18 horas às 19 horas, corresponde ao de maior transição ou maior número de locais distintos que os usuários "visitaram" ou passaram. Isso significa que os telefones celulares desses usuários ficaram acampados por pouco tempo em

[41] Poisson: A distribuição de *Poisson* é uma distribuição de probabilidade de variável aleatória discreta, que expressa a probabilidade de uma série de eventos ocorrerem em um determinado período de tempo.
[42] Função de densidade de probabilidade: Foram analisados diferentes grupos de perfil de usuários com base no número total de chamadas realizadas, considerando o mesmo intervalo de tempo. A função de densidade de probabilidade foi medida nesse cenário.
[43] Lei de potência: As leis de potência podem ser utilizadas para representar qual a probabilidade da ocorrência de eventos.
[44] Lévy Flight: O voo de Lévy é um movimento aleatório (*random walk*) baseado na função densidade de probabilidade de Lévy, proposta pelo matemático francês Paul Pierre Lévy, cujo padrão de trajetórias é caracterizado por seguidos passos de comprimentos curtos, intercalados por um passo de comprimento longo.
[45] Local físico que o usuário está mais frequentemente, ou seja, local em que o celular do usuário está regularmente acampado na antena de celular da operadora.

diferentes torres de celulares das operadoras, em função do grande volume de deslocamento ao longo do trajeto e do tempo. Esse comportamento identificado indica um padrão de comportamento que é exclusivamente dependente do tempo, o que se conclui que o indivíduo está em constante deslocamento.

Para melhor entendimento do deslocamento das pessoas em seus trajetos cotidianamente, o estudo realizado por Song et al. (2010) revelou que existe um potencial de 93% de previsibilidade média na mobilidade dos usuários, principalmente quando as pessoas se deslocam em pequenas distâncias. No entanto, para deslocamentos mais longos, a previsibilidade é mais baixa. Isto ocorre porque, em grandes trajetos, a quantidade de torres de celulares é grande e os aparelhos celulares são acampados por pouco tempo nestas torres em função de seu rápido deslocamento.

Na experiência realizada por Becker et al. (2011), houve monitoramento de 15 trajetos diferentes percorridos por carros e trem para chegarem ao mesmo destino. Nessa pesquisa, a chamada de telefone móvel permaneceu ativa em todo o percurso para gerar os registros de ligações telefônicas. Concluiu-se que é possível ser assertivo em estabelecer um padrão confiável do trajeto de uma pessoa em movimento. Esse estudo consolida a importância do uso de CDR para identificar o trajeto de pessoas praticamente em tempo real. Silva (2014) complementa o trabalho de Becker et al. (2011) quando identifica a perda de produtividade das pessoas em função do tempo gasto em seus deslocamentos no cotidiano, no trajeto pendular “casa ⇔ trabalho ⇔ casa” (Tabela 1). Esses estudos reforçam, mais uma vez, que o uso de CDR é relevante no entendimento do deslocamento urbano e da mobilidade das pessoas a partir do aparelho móvel do celular como sensor de deslocamento.

Tabela 1 - Perda de produtividade por tempo de viagem

Tempo de viagem	**Redução de Produtividade**
Até 30 minutos	0%
De 30 minutos a 1 hora	14%
De 1 hora à 1h30 minutos	16%
Mais de 1h30 minutos	21%

Fonte: Silva (2014, p. 16)

O estudo realizado por Becker et al. (2011) foi disruptivo no estudo de mobilidade urbana, porque utilizou o CDR para fornecer informações quase que em tempo real sobre a mobilidade das pessoas em larga escala e a um custo muito mais

baixo e competitivo. Pelo simples fato das pessoas carregarem o celular consigo, o aparelho acaba se tornando um sensor disponível na maior parte do tempo, já que as pessoas se deslocam de um lado para o outro ao longo de seus trajetos. No entanto, Hanson et al. (2011), em seu estudo, apresentam as limitações sobre o uso de CDR. A primeira é que o CDR somente é gerado e armazenado nos servidores das empresas de telecomunicações, quando é estabelecida uma ligação de voz ou quando uma mensagem de texto é enviada. Quando isso não ocorre, os telefones celulares são "invisíveis", ou seja, o registro de ligações – o CDR – não é gerado. A segunda limitação refere-se à identificação da posição do celular quando registrado na torre de telefonia móvel. Estudos anteriores comprovaram que há uma incerteza quanto à distância da localização dos aparelhos celulares, quando identificados pelas torres de telefonia de aproximadamente 1 mi^2 (ou 2,58 km^2).

De fato, os registros de ligações telefônicas são armazenados pelas operadoras de telefonia móvel, pois é através das ligações de voz ou mensagens de textos realizadas e recebidas que o CDR é criado. Para Song et al. (2010), as informações mais detalhadas sobre a mobilidade humana, envolvendo grande volume de pessoas, são obtidas pelos dados de CDR. Tais autores defendem o uso de CDR para os estudos de mobilidade, mas *anonimizando* esses registros para preservar a identificação dos usuários. O estudo realizado por Song et al. (2010) com uma base de dados de 10 milhões de usuários, constatou que a média de frequência de uma chamada telefônica móvel era igual a 0,50 horas.

Isaacman et al. (2010) realizaram um estudo utilizando o CDR de bilhetagem (usados para o faturamento das operadoras de telefonia) para identificar as diferenças dos padrões de mobilidade das pessoas que viviam em cidades diferentes, como por exemplo: Los Angeles e Nova Iorque. Isaacman et al. (2010) constataram que os residentes de Los Angeles percorriam uma distância de viagem casa ⇔ trabalho ⇔ casa, em média, duas vezes maior do que os nova-iorquinos. Essa informação é relevante para o contexto do planejamento e investimento mais assertivo no deslocamento das pessoas nas cidades. A técnica adotada por Marques Neto et al. (2013) para comprovar a eficácia do uso de CDR, nos eventos de larga escala como os jogos do campeonato Brasileiro de Futebol de 2011 e durante as comemorações de *réveillon* de 2011-2012, foi diferente. Coletou-se CDR do mesmo local e nos mesmos horários antes e depois do evento acontecer com o objetivo de analisar o padrão de deslocamento dos usuários ao longo de eventos de grande magnitude. A Figura 5 indica

os momentos em que foram realizadas as coletas das informações do CDR para análise do comportamento de movimento do evento de grande escala.

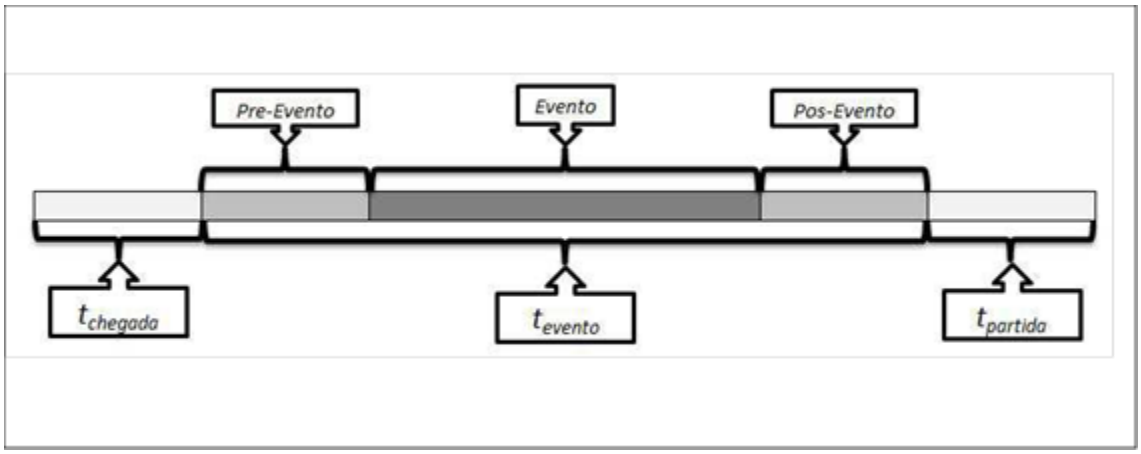


Figura 5 - Linha do tempo para a análise da carga durante um evento
Fonte: Marques Neto et al., 2013

Para González, Hidalgo, e Barabási (2008), muitos fatores ainda desconhecidos influenciam o padrão de mobilidade da população, desde meios de transportes da casa ⇔ trabalho ⇔ casa, como restrições e prioridades impostas pela condição social das pessoas. Os registros de ligações telefônicas abriram uma enorme oportunidade para correlacionar os movimentos de pessoas em grandes eventos mundiais ou, até mesmo, em surtos de doenças. Neste sentido, Isaacman et al. (2010) reforçam que compreender a mobilidade humana através do uso de CDR agiliza e contribui para o planejamento urbano das cidades e até mesmo a respostas às grandes catástrofes. Becker et al. (2011) concluíram que é possível prever o destino das pessoas, por meio de métodos precisos e cientificamente bem fundamentados, justamente porque a mobilidade diária das pessoas é, de fato, identificada por uma regularidade profundamente definida e habitual, que é caracterizada no deslocamento pendular casa ⇔ trabalho ⇔ casa. Marques Neto et al. (2013) complementam que a previsão do movimento das pessoas é importantíssima para auxiliar as empresas prestadoras de telefonia móvel a definir melhor a infraestrutura necessária para garantir o bom uso de seus serviços de telecomunicações em dias comuns e, também, em dias de grande movimento.

3.2 LIMITAÇÕES E DIFICULDADES NA ABORDAGEM

3.2.1 Questões de privacidade

Segundo Branco Junior, Machado e Monteiro (2014), a privacidade está relacionada ao interesse que as pessoas têm em manterem um espaço pessoal, sem interferências de outras pessoas ou organizações. Todavia, Faria (2010) afirma que a restrição de acesso aos dados limita, em muito, a investigação que uma comunidade de pesquisadores pode despender sobre o tema para que sejam criadas ferramentas poderosas. No que se refere ao caso da epidemia do Ebola na África Ocidental, Lenharo (2014) comenta que o grande desafio era conseguir autorização das operadoras de telefonia móvel para fornecerem informações sobre deslocamento das pessoas, pelas questões de privacidade e comerciais.

O CDR contém informações de caráter privado, como o número do telefone de quem originou a chamada (origem), do número de quem recebeu a ligação (destino), a localização[46] estimada de quem originou a chamada, entre outros importantes e sensíveis parâmetros. Essas informações são de caráter sigiloso e de responsabilidade da operadora de telefonia, exceto quando solicitada a quebra de sigilo pelos orgãos oficiais do governo, geralmente mediante mandado judicial. Segundo Faria (2010), o sigilo telefônico é garantido por lei, não somente no Brasil como também na maioria dos países do mundo. Por isso, o acesso aos dados de chamadas telefônicas é restrito.

Apesar da importância do valor do uso de CDR e da emergência ao atendimento ao surto do Ebola, Wesolowski et al. (2014) criticam a dificuldade de acesso aos registros de telefonia móvel motivadas pelas preocupações com a privacidade e questões comerciais. Mesmo em resposta a epidemias e outras emergências de saúde pública, as operadoras de telefonia não são obrigadas a fornecer acessos às suas bases de dados (CDRs).

Para Zang e Bolot (2011), apesar de insatisfatório, o processo de anonimização ainda é largamente utilizado em muitos estudos de CDR. Além disso, embora o anonimato não seja uma técnica boa e suficiente para a segurança e privacidade, os

[46] Existem diversas técnicas para que se possa localizar uma estação sem fios em uma determinada área geográfica, seja ela limitada a alguns metros quadrados ou ambientes abertos.

autores ressaltam a relevância da análise da técnica de anonimização como sendo uma excelente métrica para avaliar a vulnerabilidade da privacidade. A pesquisa de Zang e Bolot (2011), de certa forma, foi desaminadora para a liberação de dados de localização. Por outro lado, ofereceu importantes orientações sobre como os dados de localização podem ser publicados, mesmo que de maneira reduzida. A conclusão do referido trabalho é uma forte recomendação à comunidade por ser extremamente cautelosa ao publicar dados de localização de CDR anonimizados.

Outra importante consideração sobre o aspecto de privacidade é que a base de dados de CDR é a principal fonte de informação, ou o principal insumo dos sistemas financeiros no processo do ciclo da receita das operadoras, o que torna o conteúdo do CDR extremamente sensível e estratégico. Costa (2010) ressalta que todos os registros de CDR são processados e convertidos em receitas ou despesas, sejam elas voltadas para a cobrança de seus clientes ou para acertos de contas de interconexão com outras operadoras.

Outra importante consideração sobre esta questão de privacidade é referente ao Marco Civil da Internet no Brasil. Discutiu-se, durante anos, esse marco civil, que é considerado, atualmente, um texto pioneiro no mundo ao estabelecer regras e deveres no ambiente virtual. O documento é considerado a "constituição da internet". O texto aborda, entre outras coisas, a liberdade de expressão e a proteção da privacidade. O Art. 3º – a disciplina do uso da internet no Brasil – tem os seguintes princípios: (i) proteção da privacidade e a (ii) proteção dos dados pessoais.

Segundo a lei nº 12.965 (BRASIL, 2014), a guarda e a disponibilidade dos registros de conexão e de acesso a aplicações de internet de que trata a lei, bem como de dados pessoais e do conteúdo de comunicações privadas, devem atender à preservação da intimidade, da vida privada, da honra e da imagem das partes direta ou indiretamente envolvidas. Além disso, outra importante consideração é referente à proteção aos registros e aos dados pessoais, ao conteúdo das comunicações privadas – que somente poderão ser disponibilizados mediante ordem judicial, nas hipóteses e na forma que a lei estabelece.

3.2.2 Questões de processamentos computacionais e CDRs sintéticos

Grande volume de dados requer alto custo computacional para a análise e processamento. Além disso, outro ponto a ser considerado consiste no uso restrito da

massa de dados de CDR pelas operadoras de serviço de telefonia (FARIA, 2010). Por todas essas restrições expostas anteriormente (custo operacional e acesso aos CDRs), aumenta a relevância da criação do método de CDR sintético. Isaacman (2012) ressalta a importância do CDR sintético para estudos de mobilidade, o qual trouxe benefícios extraordinários. Através de CDR sintético, os pesquisadores podem explorar melhor a predição de comportamentos sociais e urbanos. Na construção do modelo de CDR sintético, é possível criar situações hipotéticas, como populações inteiras de cidades "sintéticas" (hipotéticas) e seus diferentes comportamentos de deslocamento de mobilidade urbana, apenas alterando parâmetros utilizados para a geração dos CDRs sintéticos.

Outra importante razão para o uso de CDR sintético pela comunidade científica é a preservação da identidade das pessoas. Isto porque as informações utilizadas para a geração dos campos dos CDRs sintéticos são públicas, evitando diversos problemas de privacidade associados aos usuários de telefones móveis.

É importante ressaltar que o CDR sintético é uma mímica do CDR real, construído utilizando-se dados de Censo sobre mobilidade e deslocamento das pessoas. O CDR sintético emprega cenários hipotéticos, mas sustentados sobre informações reais de pesquisas (ISAACMAN, 2012). Os CDRs sintéticos podem ser utilizados nos mesmos estudos de casos dos CDRs reais[47], inclusive, com abrangências ainda maiores, por considerar dados estatísticos de institutos de pesquisas de censos, que podem ampliar a aplicabilidade desses estudos.

Por este motivo, uma das importantes aplicações de CDRs sintéticos é no planejamento de implementação da rede, tanto para implantação quanto para o crescimento vegetativo (expansão da rede)[48]. Sua composição inclui dados estatísticos de censos de pesquisas, como a população da região, o tempo de deslocamentos de seus habitantes da casa para o trabalho, a quantidade de pessoas que trabalham no mesmo bairro em que moram. Tais dados permitem criar diferentes cenários que não podem ser

[47] Isaacman et al., 2012, p. 240: We validate our approach against large-scale location datasets drawn from two major US metropolitan areas. We compare our generated CDRs against real CDRs, and show that our location distributions achieve more than 4 times error reduction compared to a Random Waypoint model.

[48] Isaacman et al., 2012, p.240: As an example of how our models can help answer concrete questions about human mobility, we use our synthetic CDRs to compute daily ranges of travel. Our synthetic CDRs exhibit error at the median of less than 0.8 and 1 mile for NY and LA residents, respectively. This accuracy constitutes more than a 14 times improvement over that of a Weighted Random Waypoint model.

reproduzidos em ambientes reais de CDR, o que amplia, ainda mais, a importância do uso do método do CDR sintético. Esses cenários sintéticos permitem criar uma estrutura de rede de acesso para atender a demandas de crescimento imprevistas, ou, ainda, implantar a estrutura de rede de acesso móvel inicial.

4 METODOLOGIA

A metodologia do trabalho pode ser entendida como sendo a base na qual ficará assentada a pesquisa e também consiste em ser uma das principais decisões a serem inicialmente adotadas. Desta forma, seguem os critérios separados por passos[49], o método utilizado na pesquisa e os *softwares* de mercado utilizados neste projeto.

4.1 PROCESSO DE PESQUISA BIBLIOGRÁFICA

No passo primeiro, foi fundamental aprofundar as análises sobre a indústria das telecomunicações, que oferecem excelentes condições para o estudo do Big Data, em função do vasto volume de dados em tempo real, de diferentes tipos de fontes de dados não estruturados, ambiente para armazenamento de conteúdo e alta velocidade em diferentes meios de conexões de acesso as informações, tornando-se, assim, um ecossistema ideal para esta pesquisa. Ao longo desta análise, foi entendido que, para ter acesso aos registros de ligações telefônicas (CDR) para aprofundar estudos sobre telefonia seria necessário ter autorização jurídica ou ser um instituto de pesquisa com contrato de confidencialidade. Isso tudo restringia o acesso aos registros de ligações telefônicas (CDR).

No segundo passo, foi importante conhecer o trabalho idealizado pela prefeitura do Estado do Rio de Janeiro, que funciona no Centro de Operações Rio (COR), e que realizou o estudo com a UFRJ de deslocamento de pessoas no Réveillon de 2013 no Rio de Janeiro.

No terceiro passo, a partir deste encontro com os cientistas de dados da prefeitura do Rio de Janeiro, o autor desta dissertação descobriu que, para ter acesso aos registros de ligações telefônicas (CDR) a fim de realizar pesquisas, era necessário ter autorização das operadoras. Após diversas tentativas para obter os dados de CDR de uma grande operadora no Rio de Janeiro, a resposta foi sempre a mesma: que não era possível.

No quarto passo, logo em seguida, foi realizada uma nova pesquisa às bases indexadas, com as palavras-chave “CDR” e “Registro de ligações telefônicas”, levando

[49] Passos: nesse contexto, entende-se como ações realizadas e aplicadas para se chegar a formulação da tese da pesquisa nos diferentes critérios estabelecidos.

em consideração a relevância[50] dos trabalhos publicados. Como resultado, foram encontrados diversos artigos publicados utilizando CDRs reais de operadoras de diversos países.

No passo quinto, para refinar a pesquisa, foram realizadas pesquisas sobre "congressos" e "seminários" que abordassem o tema CDR. Como resultado, foi encontrada uma única conferência sobre análise de dados de telefonia móvel – Third International Conference on the Analysis of Mobile Phone Datasets –, realizada no laboratório de mídia do MIT, em Cambridge (MA), de 1º a 3 de maio de 2013. Na ocasião, o professor Dr. Sibren Isaacman apresentou um estudo sobre modelagem de mobilidade urbana – Human Mobility Modeling at Metropolitan Scales –, resultado prático de sua tese de doutorado (Modeling the impact of human mobility: mobile devices as sensors and content vectors, 2012) na Universidade de Princeton (USA). O trabalho apresentado por Isaacman introduzia a técnica do CDR Sintético[51], que reproduzia o mesmo comportamento de um CDR real.

No passo sexto, contatou-se o professor Sibren Isaacman e a ele foi relatada a dificuldade em obter CDRs reais para estudos e pesquisas. Por recomendação do professor, a sugestão foi utilizar a técnica que ele criou para gerar o próprio CDR sintético para a pesquisa. Essa técnica evita o uso de dados privados.

No passo sétimo, após leitura e entendimento da tese de doutorado do professor Sibren Isaacman, chegou-se à conclusão que a técnica desenvolvida por ele precisa ser adaptada porque, no estudo original, para comprovação do método, Isaacman utilizou como insumo os CDRs reais para comprovar a fidelidade do método de 97,5% quando comparado com estudos realizados com CDRs reais. Todavia, como a pesquisa desta dissertação era de realizar estudos com CDRs sem violar a privacidade das pessoas, não poderia ser utilizado como insumo CDRs reais. Por este motivo, essa dificuldade matemática precisava ser superada. Esse foi o desafio enfrentado ao longo desta dissertação.

[50] Relevância: segundo Heart (2011, p.39), consiste na coleta das evidências que se leva em consideração a frequência de citações no tempo.
[51] CDRs Sintéticos podem ser compreendidos como sendo CDRs hipotéticos ou CDRs artificiais.

4.2 O MÉTODO ADOTADO

Utilizou-se fonte de pesquisa de dados (IBGE, 2011), bem como modelos matemáticos e de estatística. Por esses motivos, fazer o uso das técnicas de metodologia foi mister. Segundo Silva (2005), a metodologia científica é entendida como um conjunto de etapas ordenadamente dispostas, que se deve vencer na investigação de um fenômeno. Existem várias formas de pesquisa: o método escolhido, nesta dissertação, foi laboratorial, principalmente quanto à aplicabilidade no deslocamento urbano de pessoas e na simulação de tráfego (volumetria) em redes móveis. O que caracteriza a pesquisa de laboratório é o fato de que ela ocorre em situações controladas, valendo-se de instrumental (ou método) específico e preciso. Tais pesquisas podem ser realizadas em ambientes de testes (simulação) ou em ambientes reais. A natureza desta pesquisa é de ordem prática, ampliando e viabilizando o uso do CDR. No aspecto da abordagem, a pesquisa é quantitativa, pois a técnica de CDR sintético (hipotético) pode ser (e foi, neste trabalho) perfeitamente mensurada em números e os resultados podem ser analisados utilizando-se o modelo estatístico, que foi um dos principais pilares matemáticos utilizados nesta dissertação. Segundo Dalfovo, Lana e Silveira (2008), o uso da estatística é um diferencial e tem a intenção de garantir a precisão dos trabalhos realizados, conduzindo a um resultado com poucas chances de distorções.

4.3 SOFTWARES DE MERCADO UTILIZADOS NESTE PROJETO

A seguir, são descritos os *softwares* utilizados nesta pesquisa para demonstrar o potencial da geração e, principalmente, do uso de CDR sintético.

1. *wxMaxima (www.sourceforge.net)*

Versão: 14.12.1
© 2004 - 2014 Andrej Vodopivec, Ziga Lenarcic e Doug llijev
Open Source Software.

O Maxima é um sistema de computação algébrica baseado em uma versão de 1982 do Macsyma. Ele é escrito em Common Lisp e funciona em todas as plataformas, tais como Mac OS, Linux e Microsoft Windows. Trata-se de um *software* livre, cuja licença é do tipo General Public License.

O Maxima é semelhante ao Matlab e ao Mathematica, possuindo um sistema de álgebra computacional completo e especializado em operações simbólicas. Oferece,

também, recursos numéricos, tais como integral, diferencial, sistemas de equações lineares, vetores, matrizes e aritmética de precisão arbitrária, números inteiros e racionais. O Maxima produz resultados precisos usando seu sistema especial de *floating* e pode trabalhar com funções e dados em duas ou três dimensões. É um sistema de propósito geral e cálculos de casos especiais, tais como a fatoração de números grandes e a manipulação de polinômios extremamente grandes.

O Maxima teve uma importante função neste estudo, pois, por seu intermédio, foram realizados os cálculos das duas gaussianas, gerando o gráfico e os horários das ligações realizadas pelos usuários sintéticos.

2. *OpenCellID* (www.opencellid.org)
   Open Source Software.

OpenCellID é o maior projeto de comunidade colaborativa do mundo, que coleta posições de Global Positioning System (GPS) de torres de celulares móveis, usados gratuitamente, para uma infinidade de propósitos comerciais e privados. Possui mais de 49.000 contribuintes já registrados com *OpenCellID*, contribuindo com mais de 1 milhão de novas medições todos os dias – em média – para o banco de dados do *OpenCellID*. O projeto foi criado para servir como uma fonte de dados para localização GSM. A partir de janeiro de 2015, a base de dados possuía, aproximadamente, 7 milhões de celulares GSM únicos e 1,2 bilhões de medições. O banco de dados do *OpenCellID* é publicado sob uma licença Creative Commons, de conteúdo aberto, com a intenção de promover o uso livre e redistribuição dos dados.

As identificações das torres de celulares móveis são obtidas, principalmente, pelos usuários que instalam em seus aparelhos celulares aplicativos como *OpenCellID* cliente. Estas informações são coletadas e transferidas para a base de dados do *OpenCellID*. Desta forma, é possível utilizar a base de dados do *OpenCellID* para identificar quais as prestadoras de serviços de telefonia móvel possuem cobertura em uma região geográfica específica. Os dados ainda podem identificar as células e as localizações dos aparelhos móveis, através de posição do ID do celular. A vantagem é que utilizar este serviço de localização é mais rápido do que usar sistemas de navegação baseados em satélites. Em contrapartida, é menos preciso por causa da falta de identificação de células conhecidas.

O *OpenCellID* fornece o total de células por países do mundo, o total de células por operadora de telefonia celular e o total de novas células adicionadas nas redes das operadoras por dia. Além disso, é possível visualizar as torres de celulares por região geográfica de interesse e, também, por operadora de telefonia celular, assim como gerar um mapa de calor que mostra a densidade de torres celulares cobrindo todas as regiões do globo terrestre.

O uso do *OpenCellID* nesta dissertação foi de extrema importância, pois foi possível obter a informação da posição geográfica de qualquer torre de celular de qualquer operadora em qualquer lugar do mundo. Desta forma, extraíram-se as informações de latitude e longitude de uma torre de celular de uma operadora específica e enriquecer o CDR sintético gerado neste projeto.

3. Tableau Desktop (www.tableau.com)

Versão profissional|Edition, v9.2

© 2015 Tableau Software, Inc.

O *Tableau* é um produto destinado para analisar dados e, por meio dele, é possível gerar diferentes visões gráficas para diferentes objetivos. O *Tableau Desktop* permite conectar os usuários a qualquer base de dados de tamanhos diferentes ou de qualquer formato. Por este motivo, o *Tableau* foi o *software* selecionado, entre vários outros pesquisados, para apresentar graficamente os locais (bairros) onde os usuários sintéticos realizaram ligações telefônicas. O *Tableau Desktop* tem uma abordagem totalmente nova para visualizar dados *online* e, por seu intermédio, é possível disponibilizar as visualizações em *WEB* e acessá-las de qualquer lugar e a qualquer hora.

A visualização através do *Tableau Desktop* foi construída utilizando-se o mapa do Estado do Rio de Janeiro como fundo (Figura 10) e, sobre este mapa, foram marcadas as coordenadas em latitude e longitude, indicadas com cores diferentes para cada posição (de onde cada ligação foi realizada). A indicação do usuário que realizou a ligação foi também indicada no mapa, bem como o horário e o dia que a ligação foi realizada. Desta forma, foi possível apresentar, visualmente, a localização do bairro do qual cada ligação foi realizada e os registros de CDRs sintéticos gerados através do método WHERE, demonstrado nesta dissertação. Os CDRs sintéticos gerados são muito similares aos CDRs gerados automaticamente pelas operadoras de telecomunicações.

5 FUNDAMENTAÇÃO TEÓRICA DO MÉTODO WHERE

O método ***W****ork and* ***H****ome* ***E****xtracted* ***RE****gions* (WHERE) foi criado por Isaacman (2012) para estudar o deslocamento de grande quantidade de pessoas dentro de diferentes áreas de uma cidade, podendo ser implementado utilizando-se, exclusivamente, distribuições de probabilidades temporais e espaciais para a produção de CDR sintético, gerado a partir de dados públicos de institutos ou censos de pesquisas.

Para a comprovação da teoria, foram empregados registros de CDRs reais de telefonia móvel no batimento dos resultados finais contra os resultados obtidos por meio de pesquisas com dados públicos (censos geográficos, por exemplo), conforme recomendação original do método WHERE. Essa pesquisa, realizada por Issacman et al. (2012), obteve o suporte da Fundação Nacional de Ciência dos EUA, da Universidade de Engenharia Princeton e de pesquisadores do laboratório da AT&T[52] e utilizou como fonte de dados CDRs de centenas de milhões de telefones móveis das áreas metropolitanas das cidades de New York e Los Angeles.

Segundo Isaacman et al. (2012), a modelagem de distribuições espaciais e temporais de probabilidades de dados sobre a população tem como objetivo gerar sequências de locais e de horários em que, supostamente, ligações teriam sido realizadas para qualquer número sintético de pessoas, em qualquer região dentro do escopo espacial considerado. Como já mencionado, esse resultado pode ser obtido mediante resultados de pesquisas de censo e de distribuições temporais de ligações telefônicas ao longo do dia, disponíveis em trabalhos já publicados.

Outro benefício do modelo WHERE é permitir mais flexibilidade, podendo definir cenários específicos de testes, além de ser mais compacto (aproximadamente 2 gigabytes), se comparado aos dados reais de CDR (em torno de 100 gigabytes)[53]. Além disso, o modelo pode ser disponibilizado e utilizado pela comunidade de pesquisadores por não reproduzir nenhum padrão de mobilidade de qualquer pessoa real, evitando, assim, muitas das preocupações de privacidade associadas ao CDR real, em geral,

[52] Isaacman et al., 2012, p.251: We thank our shepherd, Rajesh Balan, and the anonymous reviewers for their feedback. Parts of this work were supported by the National Science Foundation under Grant Nos. CNS- 0614949, CNS-0627650, and CNS-0916246. Isaacman also acknowledges support from a Princeton University Wallace Memorial Fellowship in Engineering and a research internship from AT&T Labs.

[53] Isaacman et al., 2012, p.240: Second, our model for a metropolitan area with a 50-mile radius can be stored as a set of histograms that fit within 2 gigabytes. In contrast, an anonymized CDR dataset for the same area occupied approximately 100 gigabytes.

obtido nas operadoras de telefonia. Outro desafio, conforme já foi explicado anteriormente, é obter a autorização das empresas de telefonia para manipular os CDRs reais para pesquisas.

Os CDRs sintéticos (hipotéticos) têm o mesmo formato e são modelados para aproximar os padrões de movimento das pessoas, deduzidos a partir dos CDRs reais. O aumento da complexidade do modelo dos CDRs sintéticos (hipotéticos) resultará em padrões de movimento mais precisos, que, por sua vez, produzirá maior fidelidade dos CDRs sintéticos (hipotéticos) se comparados aos CDRs reais.

A mobilidade urbana está intimamente ligada à geografia das cidades onde as pessoas vivem. Portanto, qualquer modelo de mobilidade preciso deveria levar em consideração tanto particularidades da área geográfica tratada quanto o padrão de mobilidade das pessoas.

Os dados de entrada para a construção de CDRs sintéticos são dados estatísticos oriundos de pesquisas de censo publicadas oficialmente. Esses dados aproximam o modelo do cenário de estudo de populações reais e de movimentos dentro das cidades[54].

A Figura 6 apresenta uma visão geral das fontes de dados obtidos de pesquisa de censo utilizados na produção do CDR sintético.

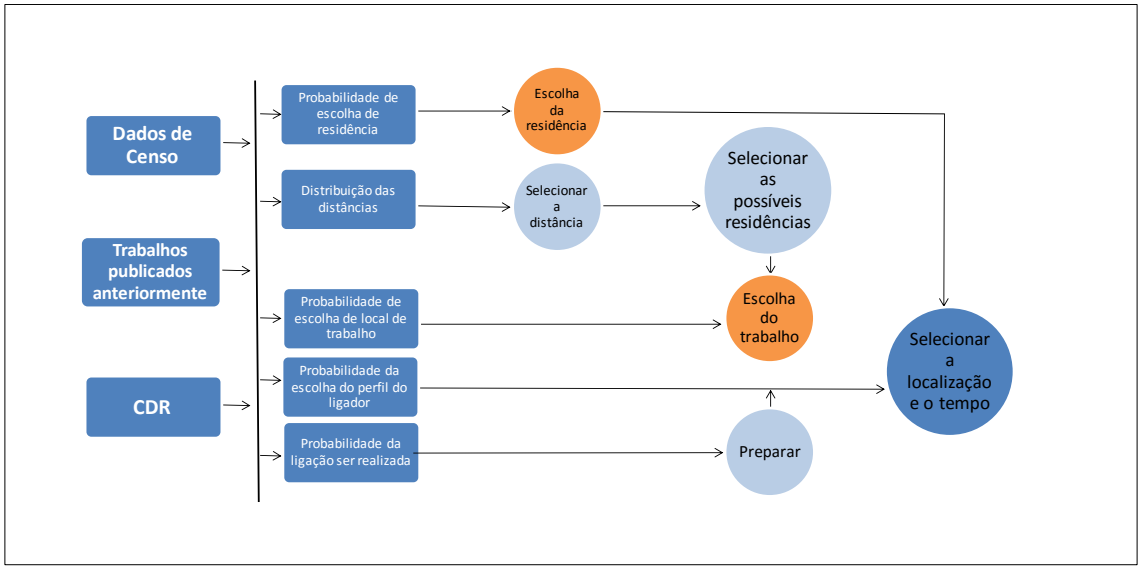


Figura 6 - Visão geral da abordagem de modelagem WHERE
Fonte: Isaacman et al., 2012, p.241

[54] Isaacman et al., 2012, p.240: The technique is extensible to greater levels of precision by providing it more complete input probability distributions (at the cost of increased model complexity).

Distribuição	Fontes de entrada de dados		
	Pública	Híbrido	Todos CDR
Residência	Censo	Censo	CDR
Distribuição de distância	Censo	Censo	CDR
Trabalho	Censo	Censo	CDR
Hora em hora	Censo residência e trabalho por hora do dia	CDR	CDR
CallTime e PerUserCallsPerDay	Referência de trabalhos publicados[55] [56]	CDR	CDR

Quadro 5 - Distribuições de probabilidades utilizadas no método WHERE
Fonte: Isaacman et al., 2012, p.241

O Quadro 5 apresenta as cinco distribuições de probabilidades utilizadas no método WHERE. Os dados de entrada para essas probabilidades podem ser recolhidos a partir de diferentes fontes, podendo o método ser baseado inteiramente em dados públicos, ao contrário dos métodos que utilizam dados reais, que são proprietários e exclusivos das operadoras (ISAACMAN et al., 2012).

O método compreende duas fases:

1) *Create* – em que são sorteados os usuários, seus locais de trabalho e residência, assim como outros atributos ligados à quantidade de ligações feitas em determinado dia (que no método adaptado apresentado nesta dissertação será chamado de "perfil") – ver Quadro 6.

[55] ISAACMAN, S. et al. (2012, p.251): ALMEIDA, S.; QUEIJO, J.; CORREIA, L. Spatial and temporal traffic distribution models for GSM. In: Vehicular Technology Conference, Sept. 1999.
[56] ISAACMAN, S. et al. (2012, p. 251): J. Candia, M. C. González, P. Wang, T. Schoenharl,G.Madey, and A.-L. Barabási. Uncovering individual and collective human dynamics from mobile phone records. MATH.THEOR., 41:224015, 2008.

Algorithm 1 Create

Ensure: $pop[]$ is an N sized array of 4-element structures to be filled in with 4 properties for each of N synthetic users

1: **for** $user = 0 \rightarrow N$ **do**
2: $pop[user].home \leftarrow$ location from *Home*
3: $commute \leftarrow$ distance from *CommuteDistance* conditioned on $pop[user].home$
4: $pop[user].work \leftarrow$ location from *Work* at distance $commute$ from $pop[user].home$
5: $pop[user].callsbehavior \leftarrow$ user type from *CallTime*
6: $pop[user].callsperday \leftarrow \mu$ and σ from *PerUserCallsPerDay* and independent of $pop[user].callsbehavior$
7: **end for**

Quadro 6 – Algoritmo *Create*
Fonte: Isaacman et al., 2012[57]

Abaixo é descrito o algoritmo Create.

pop[] é uma matriz de estruturas de 4 elementos a ser preenchida com 4 propriedades para cada um dos ***N*** usuários

1: Loop com ***usuário*** de zero a ***N***, faça

2: ***pop[usuário].casa*** recebe a localização da variável ***Casa***

3: ***deslocamento*** recebe ***DistanciaDeslocamento*** de acordo com ***pop[usuário].casa***

4:***pop[usuário].trabalho*** recebe localização do ***Trabalho*** a uma ***DistanciaDeslocamento*** a partir de ***pop[usuário].casa***

5: ***pop[usuário].comportamentochamadas*** recebe tipo de usuário de ***HoraChamada***

6: ***pop[usuário].chamadaspordia*** recebe μ e σ retirados de ***ChadamasUsuarioPorDia*** e independentemente de ***pop[usuário].comportamentochamadas*** .

7 :fim do loop

[57] Isaacman et al. (2012, p.243): Distribution of call times for two classes of users as determined by X-means clustering.

2) *Move* – onde são sorteadas as quantidades de ligações e seus horários para cada dia da semana e associado a cada um destes resultados o local onde o usuário supostamente[58] se encontra no intervalo de tempo da respectiva ligação – ver Quadro 7.

Algorithm 2 Move

1: **for** $user = 0 \rightarrow N$ **do**
2: **for** $day = 0 \rightarrow D$ **do**
3: $callstoday \leftarrow$ normal random number with μ and σ from $pop[user].callsperday$ distribution
4: **for** $call = 0 \rightarrow callstoday$ **do**
5: $calltime \leftarrow$ time from $pop[user].callsbehavior$
6: $location \leftarrow$ location using probabilities of $pop[user].work$ and $pop[user].home$ at time $calltime$ from the *Hourly*
7: **print** $user, day, calltime, location$
8: **end for**
9: **end for**
10: **end for**

Quadro 7 – Algoritmo *Move*
Fonte: Isaacman et al., 2012[50]

Abaixo é descrito o algoritmo Move.

1: Loop com ***usuário*** de zero a ***N***, faça

2: Loop com ***dia*** de zero a ***D***, faça

3: ***chamadashoje*** recebe distribuição normal aleatória com μ e σ retirados de ***pop[usuário].chamadaspordia***

4: Loop com ***chamada*** de zero ***chamadashoje*** faça

5: horachamada recebe hora de ***pop[usuário].comportamentochamadas***

6: ***localização*** recebe localização usando probabilidades de ***pop[usuário].casa*** e ***pop[usuário].trabalho*** no instante ***horachamada*** da variável ***DeHoraEmHora .***

7: Imprima ***usuário***, ***dia***, ***horachamada***, ***localização***

8: fim do loop

[58] Supostamente é o termo utilizado para dizer que a localização exata do usuário não pode ser afirmada devido ao erro de precisão da localização geográfica da torre de telefonia celular. Alguns estudos informam que a diferença em metros pode chegar de 100 a 300 metros do posicionamento real.

9: fim do loop

10: fim do loop

Para que se obtenha um resultado mais realista com o algoritmo *Move*, já que não há CDRs reais neste estudo, nem estatísticas baseadas nesses registros, é necessário um modelo matemático que permita sortear horários de ligações ao longo de um dia, de tal forma que o total de ligações realizadas reflita o que ocorre em um centro urbano. Para isso, o método descrito por Isaacman (2012) recomenda o trabalho de Queijo e Almeida (1998).

Queijo e Almeida (1998), em seu trabalho sobre modelos para distribuições espaciais e temporais de tráfego em Global System for Mobile (GSM), estabeleceram que a densidade de tráfego é dada pelo modelo de variação espacial (amplitude) e a forma é dada pelo modelo temporal de duas gaussianas.

Para este trabalho, seguindo a recomendação de Isaacman (2012) para modelar o ritmo das ligações de cada usuário ao longo de um dia, utilizou-se a dupla gaussiana utilizada no trabalho de Queijo e Almeida (1998) – ver Gráfico 2.

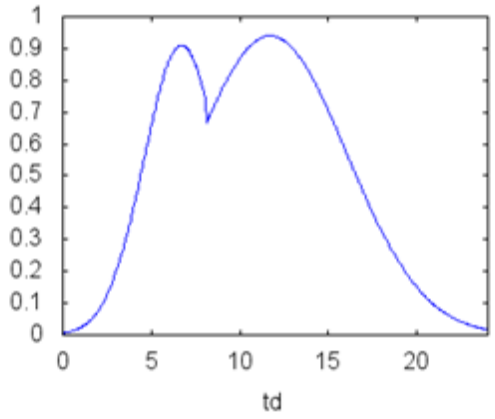


Gráfico 2 - Dupla gaussiana
Fonte: Elaborado pelo autor

Os horários das ligações de cada usuário gerado precisam ser sorteados de tal forma que seu histograma se ajuste à curva de probabilidade que define o padrão de ligações para aquela população.

O processo para realizar este sorteio compreende, segundo Olver e Towsend (2006), as seguintes fases:

(i) Definir função de distribuição de probabilidades;

(ii) Criar função de distribuição acumulada de probabilidade;

(iii) Sortear um valor entre zero e um na imagem da distribuição acumulada de probabilidade e descobrir qual o seu correspondente entre os possíveis valores da variável aleatória que está sendo considerada.

No caso de ligações ao longo de um período de 24 horas, poder-se-ia exemplificar o método com a função abaixo, que infere que uma pessoa faz, em média, no período de 6h às 18h, cinco vezes mais ligações por hora do que no período de zero hora até às 6h; e, no período de 18h às 0h, duas vezes mais ligações por hora do que no período de zero hora até às 6h. Uma possível função que atenda a esses requisitos é a função degrau, exibida no Gráfico 3.

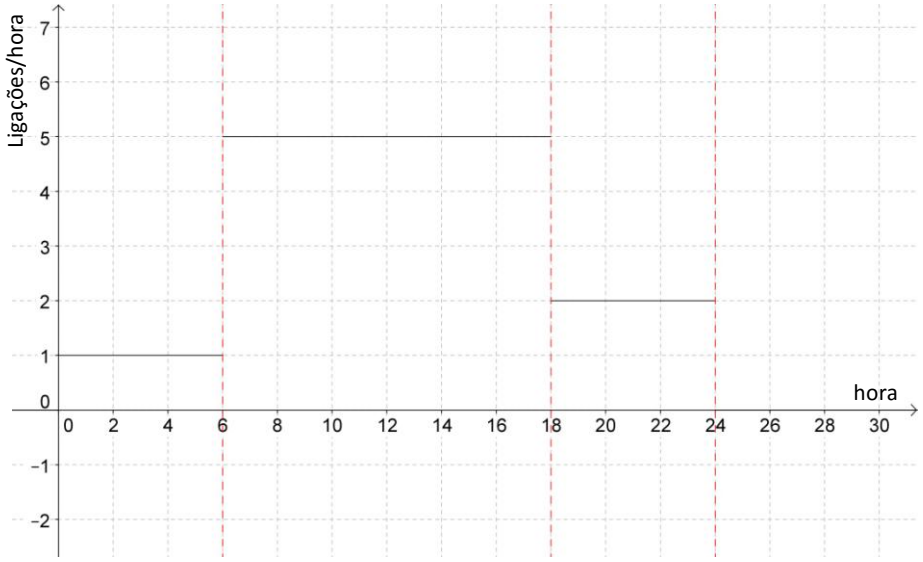


Gráfico 3 – Variação da frequência das ligações ao longo de um dia
Fonte: Elaborado pelo autor

A área entre o gráfico e o eixo horizontal é igual a 78. Dividindo todas as cotas verticais por 78, chega-se ao gráfico normalizado abaixo (Gráfico 4) que nada mais é do que uma função densidade de probabilidade, uma vez que a área entre o gráfico e o eixo horizontal vale 1.

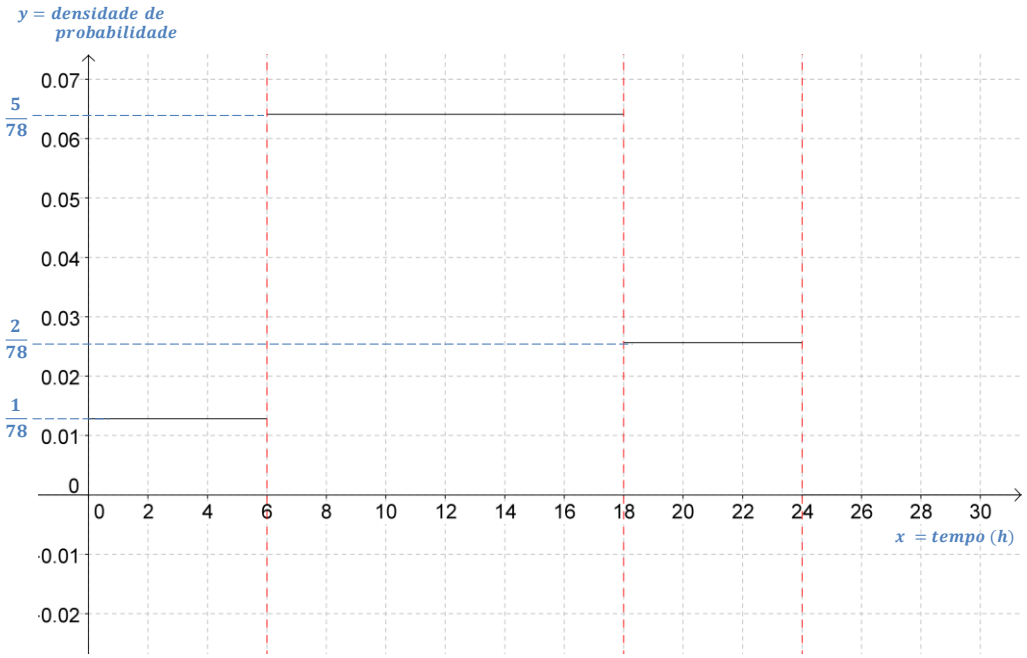


Gráfico 4 - Função densidade de probabilidade
Fonte: Elaborado pelo autor

Note que a área total sob a curva vale 1, assim como ocorre em uma gaussiana.

Em seguida, foi construída a função acumulada de probabilidade dessa distribuição (Gráfico 5).

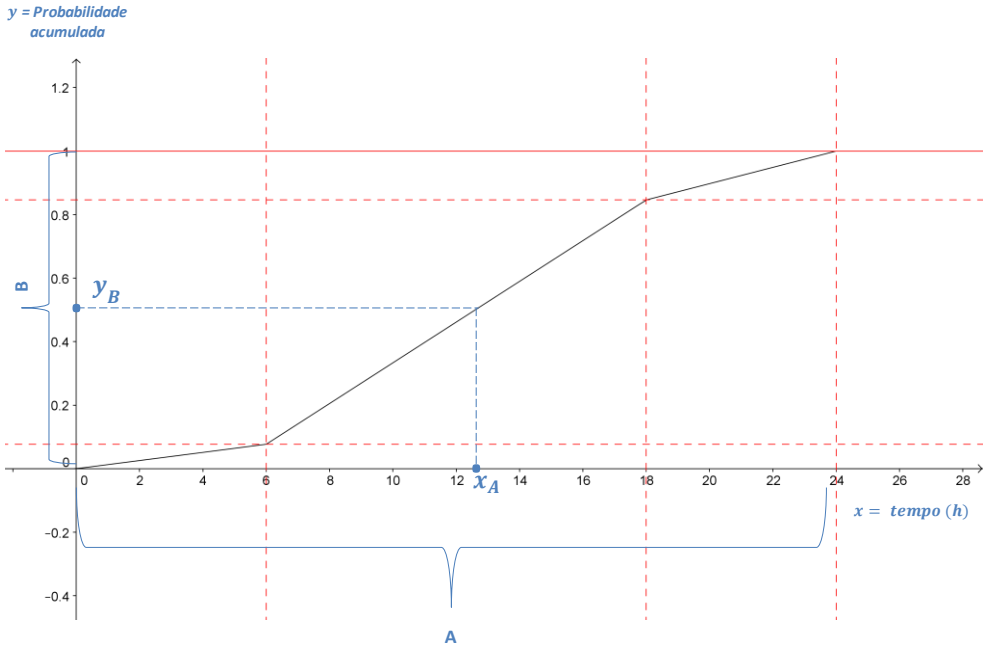


Gráfico 5 - função de distribuição acumulada de probabilidade
Fonte: Elaborado pelo autor

O próximo passo será sortear um valor qualquer em B (y_B) e descobrir qual o valor em A (x_A) dá como imagem y_B.

Para exemplificar a eficácia do método, foram gerados 200 valores aleatórios para y_B com o *software Máxima*. Foram obtidos os seus respectivos valores x_B e, assim, foi possível montar o histograma representado no Gráfico 6.

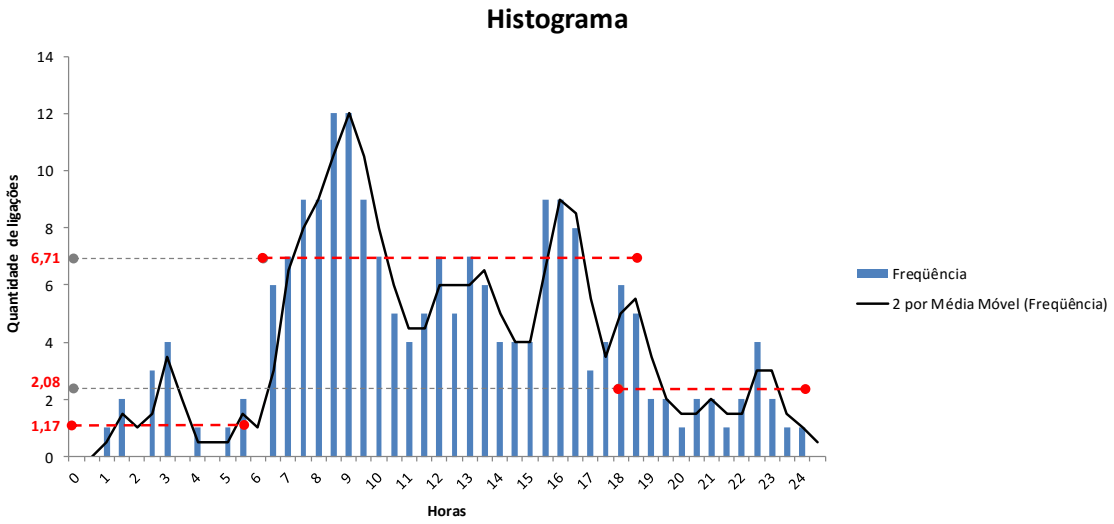


Gráfico 6- Ligações sorteadas segundo função degrau
Fonte: Elaborado pelo autor

Os pontos indicados em vermelho, no eixo vertical do Gráfico 6, representam a média de ligações a cada meia hora, nos períodos segregados pelo gráfico em degraus – *variação da frequência das ligações ao longo de um dia*. Desta forma, no período de zero hora até às 6h, houve, em média, 1,17 ligações a cada meia hora. Esse mesmo raciocínio aplica-se aos períodos segregados em vermelho: das 6h até às 18h, das 18h até às 20h e das 20h às 0h. Pode-se notar que as proporções entre as médias das ligações dos três períodos refletem com boa aproximação aquelas que caracterizam o gráfico em degraus – *variação da frequência das ligações ao longo de um dia*. Para facilitar a visualização, foi traçada uma curva de tendência, contornando o histograma. A lista completa dos horários de ligações sorteados está no Apêndice V.

Aplicando o mesmo método à dupla gaussiana, chega-se ao histograma representado no Gráfico 7.

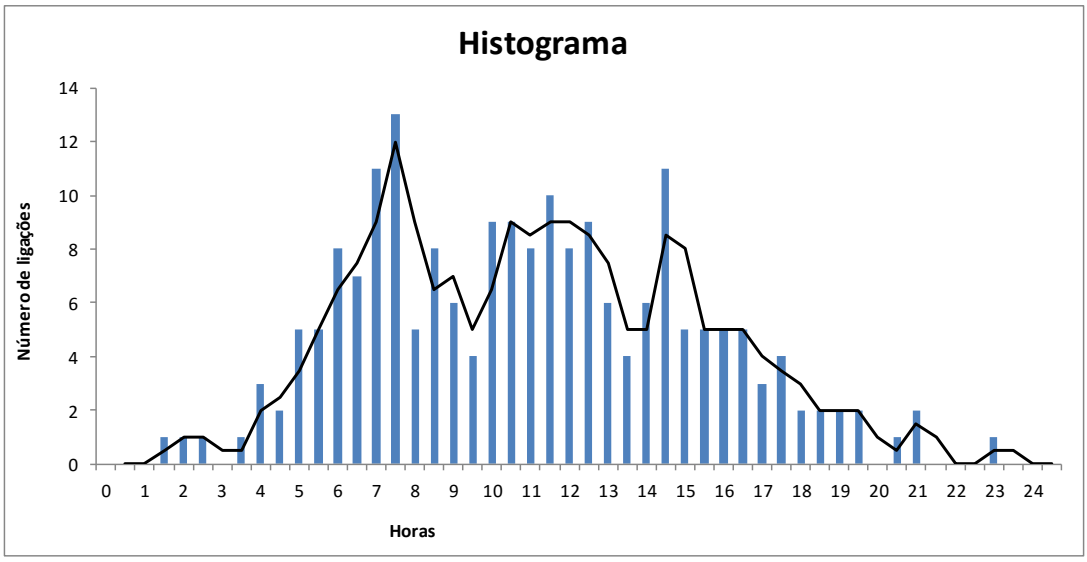


Gráfico 7 - Ligações sorteadas segundo uma dupla gaussiana
Fonte: Elaborado pelo autor

O Gráfico 7, obtido a partir de uma amostra de 200 sorteios com o método descrito em Olver e Towsend (2006), mostra que os picos da primeira e da segunda ondas da dupla gaussiana ocorrerão, respectivamente, como determina o modelo apresentado no Gráfico 2, referente à dupla gaussiana, por volta das 6h e 11h. Estes valores são "valores desviados", como será visto mais adiante, que correspondem, respectivamente, aos horários físicos em torno das 11h e das 16h.

Segundo Queijo e Almeida (1998), os sistemas celulares apresentam tráfego com características distintas do tráfego de redes fixas. Contudo, a teoria desenvolvida para a rede fixa pode ser aplicada com uma boa aproximação ao tráfego de redes celulares, com o tráfego medido em *Erlang*. Um *Erlang* corresponde a um canal (circuito) ocupado durante uma hora. Pode-se calcular o tráfego médio de um sistema A mediante a expressão indicada no Quadro 8. A lista completa dos horários de ligações sorteados está no Apêndice VI.

$$A_{[Erl]} = \frac{N_{ch} / h \times \overline{T_{chamada}}}{3600\ s}$$

Onde:

N_{ch} / h = *é o número de chamadas efetuado por hora*

$\overline{T_{chamada}}$ = *é a duração média de uma chamada*

Quadro 8 - Tráfego é medido de um sistema A, em Erlang
Fonte: Queijo e Almeida, 1998

O tráfego médio da fórmula de *Erlang* é o número de chamadas por hora multiplicada pela fração de hora que representa o tempo médio das chamadas. Ou seja, se houve três chamadas em uma hora e elas demoraram em média 20 minutos (1/3 de hora), o valor do tráfego médio será 3 x (1/3) = 1 chamada por hora.

Ainda, de acordo com tais autores, dado o tráfego oferecido a uma rede e o seu número de canais/troncos N, é possível determinar a probabilidade de um utilizador não conseguir acessar a rede, probabilidade de bloqueio, Pb, por meio da fórmula de "*Erlang* B" (para maior aprofundamento desse conceito, ver Anexo I) representada pela equação indicada no Quadro 9.

$$\boldsymbol{B}(\kappa, \lambda) = \frac{\boldsymbol{\lambda}^{\kappa}}{(\kappa!) \sum_{i=0}^{k} \frac{\boldsymbol{\lambda}^{i}}{i!}}$$

B é a probabilidade de que chamadas sejam bloqueadas por ausência de troncos livres, dentro de uma unidade de tempo, quando a intensidade média de tráfego é de **λ** chamadas por unidade de tempo, e existem k troncos para servir às chamadas.

Quadro 9 - Probabilidade de um utilizador não conseguir acessar a rede
Fonte: Queijo e Almeida, 1998

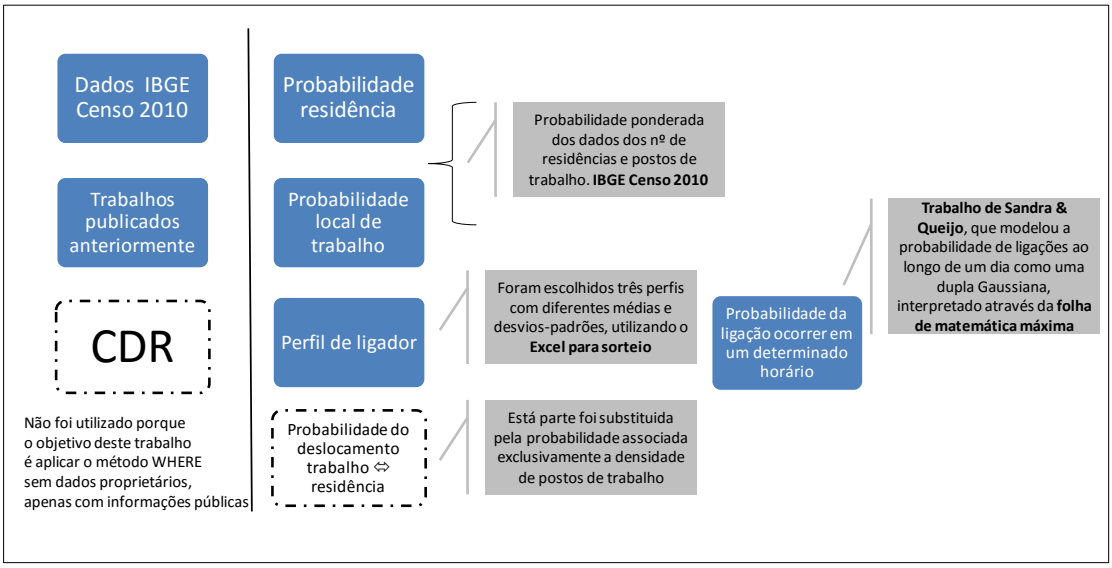
6 GERANDO CDR SINTÉTICO

Será usado um modelo simplificado – adaptado para melhorar didaticamente a abordagem do assunto – para ilustrar a aplicação do método desenvolvido pelo professor Dr. Sibren Isaacman, do Departamento da Ciência de Computação da Universidade Loyola Maryland nos EUA. A simplificação visa a dar mais importância ao método do que às dificuldades computacionais envolvidas, até porque estas dificuldades técnicas já estão superadas pelos *softwares* de mercado. No caso de uma distribuição geográfica, por exemplo, as regiões geográficas cujos dados são fornecidos pelo censo (realizadas por institutos de pesquisas) dificilmente serão quadradas, mas, quaisquer que sejam as regiões consideradas, uma possível maneira de modelar esta distribuição é quadriculando o mapa e criando uma regra para decidir qual dado será utilizado nas bordas das regiões.

6.1 GERAÇÃO DE CDR SINTÉTICO ADAPTADO

A primeira adaptação foi a opção de escolha de não se utilizar o CDR como dado de entrada porque o objetivo principal deste trabalho é justamente a produção sintética do CDR. Por este motivo, as informações sobre o número de pessoas residentes por bairro e postos de trabalho foram obtidas no IBGE – Censo 2010. Como segunda adaptação ao método WHERE, diferente do estudo original, não foi estabelecida uma distância média de deslocamento das pessoas do trajeto casa-trabalho, trabalho-casa. Nesta pesquisa, optou-se por realizar um sorteio das pessoas residentes em um determinado bairro, utilizando-se a densidade calculada a partir das informações obtidas no IBGE – Censo 2010. Para um melhor ajuste, pode-se realizar um sorteio do tempo de deslocamento das pessoas de suas residências aos seus trabalhos, utilizando dados de outras fontes públicas de tempo de deslocamento entre o trajeto casa-trabalho e trabalho-casa. A terceira adaptação realizada foi em relação à montagem do perfil de ligador. A decisão foi utilizar três perfis com diferentes médias e desvio-padrão, utilizando o sorteio do Excel. A quarta adaptação ao método original foi utilizar o estudo realizado por Olver e Townsend (2006), que orienta como realizar sorteios sobre um espaço não equiprovável modelado por uma função matemática qualquer com variáveis aleatórias comuns, através de inversão de função de probabilidade acumulada. No Quadro 10, é apresentado, em módulos, o método WHERE adaptado.

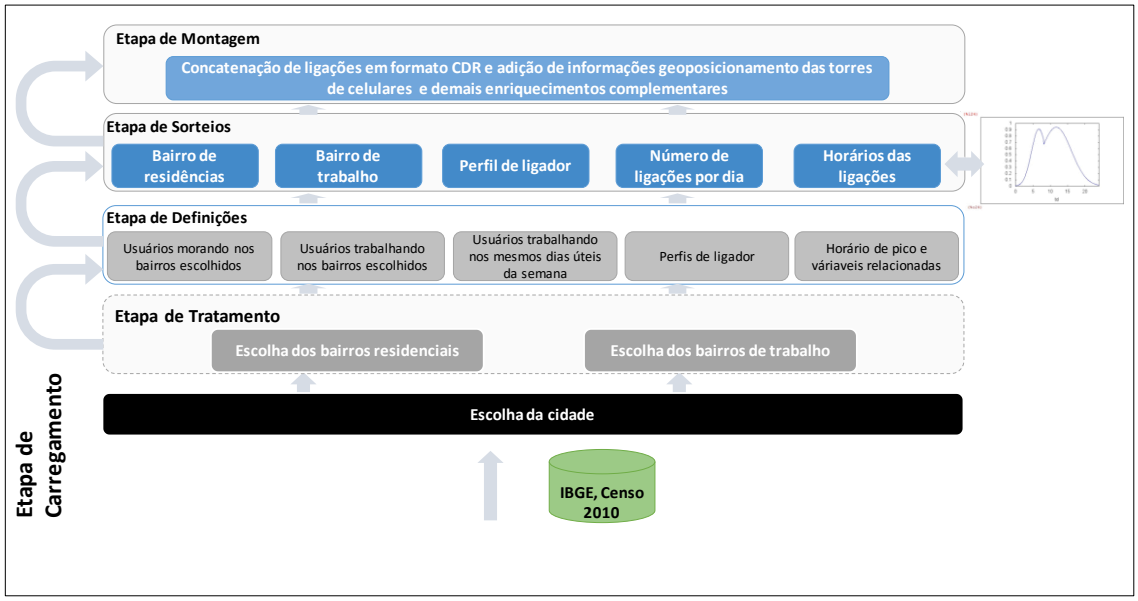
O método ***W**ork and **H**ome **E**xtracted **RE**gions* (WHERE) adaptado, descrito a seguir, foi desenvolvido e modelado pelo autor desta dissertação, com o objetivo de resolver o problema da falta de variação temporal de frequência de ligações, que determina o padrão de um determinada população, tendo em vista a ausência de uso de CDRs reais.



Quadro 10 - Método WHERE adaptado
Fonte: Elaborado pelo autor

A estrutura (o *framework*) – Quadro 11 – do método WHERE adaptado constou de cinco etapas para a produção final do CDR sintético, conforme descrito a seguir:

(i) A primeira etapa, chamada de etapa de carregamento, tem como objetivo selecionar os dados do censo relativos às cidades escolhidas, bem como definir quantidades totais de usuários;
(ii) A segunda etapa, chamada de etapa de tratamento, processa a escolha dos bairros de residência e de postos de trabalho;
(iii) A terceira etapa, chamada de etapa de definições, convenciona que os usuários moram e trabalham nos bairros escolhidos e com horário de trabalho das 9 às 18 horas. Calcula o número de pessoas que moram e trabalham em cada bairro, definem os horários de picos das ligações, assim como os perfis de ligadores que serão atribuídos aos usuários;
(iv) A quarta etapa, chamada de etapa de sorteios, sorteia para cada usuário sintético o bairro para sua residência, o local de trabalho, o perfil de ligador, a quantidade de ligações realizadas por dia e os horários em que as ligações ocorrem;
(v) A quinta etapa, chamada de etapa de montagem, é a que concatena as ligações em formato de CDR e adiciona informações de geoposicionamento das torres de celulares e demais enriquecimentos para tornar os cenários sintéticos (hipotéticos) mais próximos do ambiente real.



Quadro 11 - Framework do método CDR sintético
Fonte: Elaborado pelo autor

6.2 IMPLEMENTAÇÃO DO MÉTODO CDR SINTÉTICO

Seguindo a metodologia da seção anterior – Metodologia de geração de CDR sintético adaptada –, serão utilizadas, neste estudo, pesquisas realizadas por institutos de pesquisas e órgãos oficiais do governo, a saber, do Censo Demográfico de 2010, realizado pelo IBGE. Neste trabalho, as informações demográficas são, de fato, reais.

Em lugar de quadriculado geográfico, o que turvaria a visão do método, por conta de detalhes geográficos, optou-se por utilizar unidades maiores, no caso, bairros, não dando importância às suas formas originais ou tamanhos relativos. Portanto, são apresentados os mapas dos bairros da cidade do Rio de Janeiro (Figura 7) para o exercício prático do método WHERE.

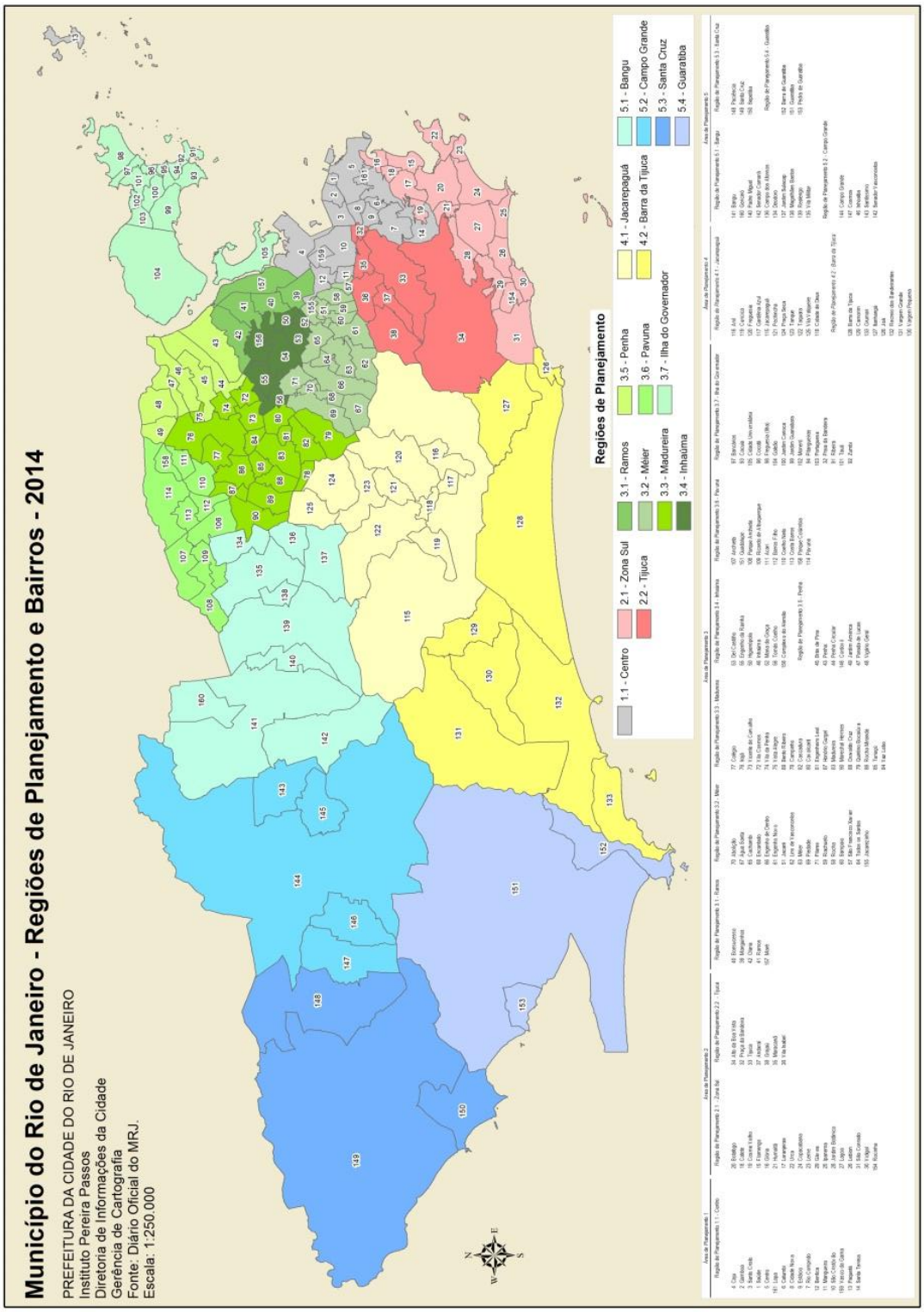


Figura 7 – Mapa da cidade do Rio de Janeiro
Fonte: Prefeitura da Cidade do Rio de Janeiro, Instituto Pereira Passos, Gerência de Cartografia [59]

[59] Disponível em: <http://www.armazemdedados.rio.rj.gov.br/arquivos/1314_bairros%20-%202014.JPG>. Acesso em: 06 abr. 2015.

Além das questões de privacidade, a utilização de dados reais (do censo) faz com que cenários reais possam ser recriados ou criados para ensaios específicos. Segundo Isaacman (2012), os ensaios realizados em seu estudo possuíram uma taxa de assertividade do método WHERE de 97,5%[60].

O método WHERE foi construído utilizando-se o deslocamento das pessoas de casa para o trabalho, lugares onde mais facilmente é possível identificar o posicionamento das pessoas ao longo do dia (ISAACMAN et al., 2012).

O primeiro passo é sortear onde as pessoas residem. A cidade do Rio de Janeiro foi escolhida como modelo para esta prática da construção do CDR sintético. De acordo com Mihessen, Machado e Pero (2014), a região metropolitana do Rio de Janeiro tem o maior tempo médio de deslocamento de casa ao trabalho se comparada a outras cidades do país. Para não previlegiar nenhum bairro específico do Rio de Janeiro – e para que essa escolha não influencie voluntariamente a geração do CDR sintético (hipotéticos) –, os bairros aleatoriamente escolhidos do município do Rio de Janeiro, que para fazem parte da construção do CDR sintético, foram: (i) Barra da Tijuca, (ii) Botafogo, (iii) Centro, (iv) Copacabana, (v) Madureira, (vi) Meier, (vii) São Cristóvão e (viii) Tijuca – Figura 8.

[60] Isaacman et al. (2012, p. 50): The results lead to three conclusions. First, the algorithm is generally successful; it finds important locations for 97.5% of the user population.

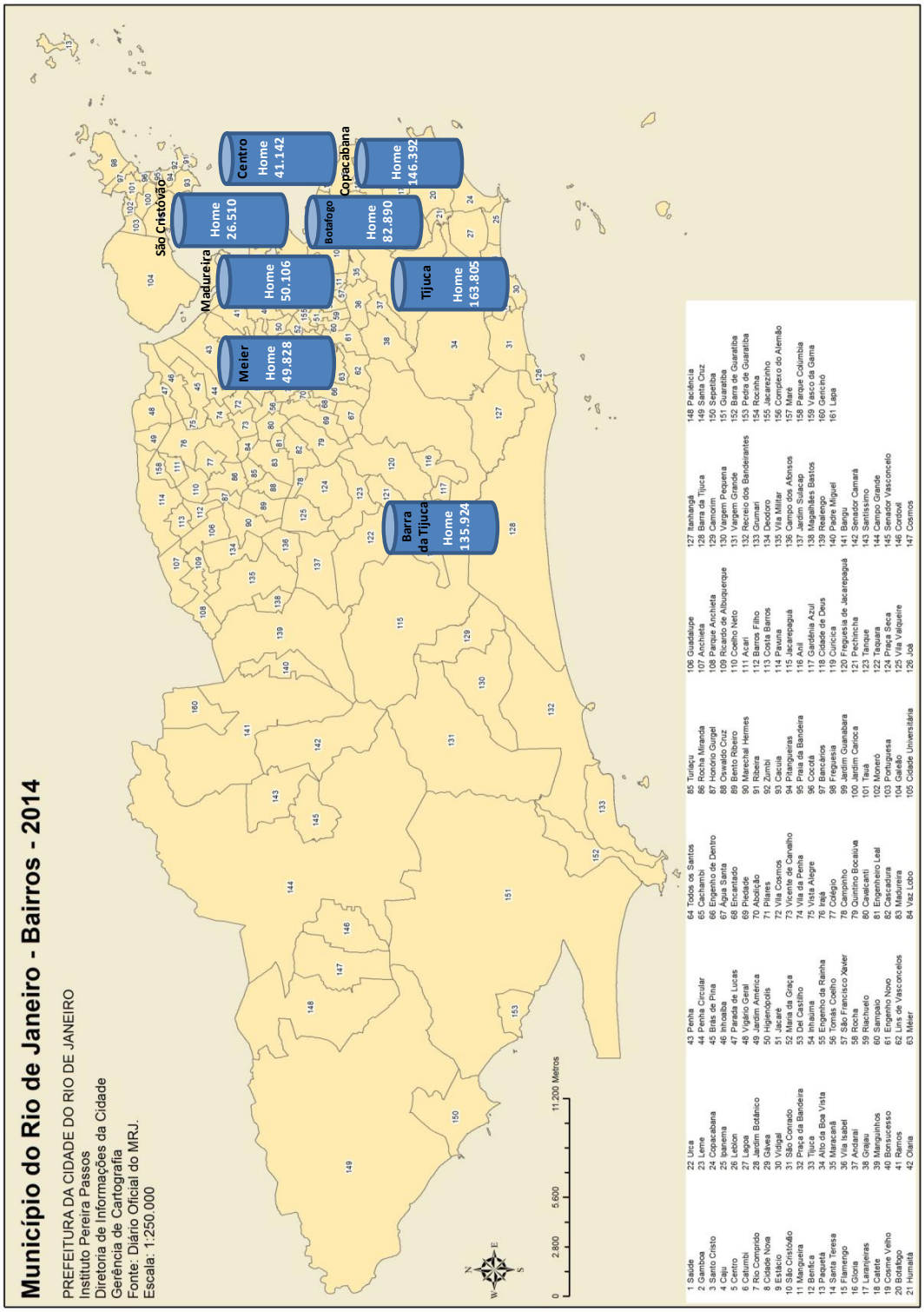


Figura 8 - População residente por bairro do Município do Rio de Janeiro
Fonte: IBGE – Censo 2010

Segundo Isaacman et al. (2012), deve-se escolher a localização da residência pela distribuição geográfica de probabilidade[61] em função da quantidade de residentes do bairro. O Quadro 12 apresenta o total de residentes por bairro em cada um desses lugares, totalizando 696.597 residentes. Segundo Isaacman (2012), os números de residências e de postos de trabalho das pessoas, em particular, são facilmente encontrados publicamente em censo geográfico de pesquisas[62].

Bairros	Residentes
Barra da Tijuca	135.924
Botafogo	82.890
Centro	41.142
Copacabana	146.392
Madureira	50.106
Meier	49.828
São Cristóvão	26.510
Tijuca	163.805

Quadro 12 - Total de população residente por bairro do Município do Rio de Janeiro
Fonte: IBGE – Censo 2010

Percentualmente, ao sortear um habitante qualquer que more na região isolada (conjunto dos bairros escolhidos para sortear os habitantes), a probabilidade dele morar num dos bairros selecionados será dada pela probabilidade descrita na Tabela 2.

Tabela 2 - Distribuição de probabilidade de um habitante que more na região isolada[63] residir nestes bairros

Bairros	Residentes
Barra da Tijuca	20%
Botafogo	12%
Centro	6%
Copacabana	21%
Madureira	7%
Meier	7%
São Cristóvão	4%
Tijuca	23%
Total	100%

Fonte: Elaborado pelo autor

[61] Isaacman et al. (2012, p. 241): First, we pull home locations randomly from a probability distribution across latitude and longitude expressing the likelihoods of where people live, i.e. Home.

[62] Isaacman (2012: p. 71): The number of people living or working in a particular area is easily found in the census and therefore is a matter of public record.

[63] Pode-se entender pela palavra “isolada” o significado de selecionada ou escolhida.

O processo de sorteio pode ser constituído de maneira bem simples[64]. Considerando cada bairro uma célula, valores são atribuídos às células de acordo com a porcentagem assinalada, como, por exemplo, os exibidos no Quadro 13.

Bairros	Residentes
Barra da Tijuca	1 a 20
Botafogo	21 a 32
Centro	33 a 38
Copacabana	39 a 59
Madureira	60 a 66
Meier	67 a 73
São Cristóvão	74 a 77
Tijuca	78 a 100

Quadro 13 - Quantidade de pessoas que podem residir nestes bairros
Fonte: Elaborado pelo autor

Para sortear uma pessoa e identificar onde ela reside, em um dos bairros mencionados no Quadro 13, será utilizada a função randômica do Excel apresentada no Quadro 14.

f(x) = TRUNCAR(ALEATÓRIO()*100)+1

Onde:

f(x) = É o resultado aleatório relacionado ao bairro onde a pessoa reside.

Quadro 14 - Escolha aleatória para identificar residência de indivíduo
Fonte: Elaborado pelo autor

Os primeiros cinco números sorteados foram: 19, 35, 77, 86 e 67.

Em função do resultado do sorteio, será posicionado o número sorteado nos bairros escolhidos: (i) Barra da Tijuca, (ii) Botafogo, (iii) Centro, (iv) Copacabana, (v) Madureira, (vi) Meier), (vii) São Cristóvão ou (viii) Tijuca.

[64] Esta é apenas uma sugestão escolhida de forma a tornar a abordagem didática mais eficiente. Na prática, qualquer forma de sorteio que levar em consideração as probabilidades associadas a cada bairro servirá.

A função representada no Quadro 14 trará valores que irão de 1 a 100. Se o número que sair for 76, saber-se-á que a pessoa é moradora de São Cristóvão. Caso o número sorteado seja 2, o morador será do bairro Barra da Tijuca e, assim, sucessivamente.

Desta forma, se forem sorteadas, por exemplo, 200 pessoas, será possível definir, exatamente, onde cada a pessoa reside e, por conseguinte, ter-se-á uma lista sequencial dos registros, conforme descrito abaixo:

- (1º usuário, Barra da Tijuca,...)
- (2º usuário, Centro,...)
- (3º usuário, São Cristóvão,...)
- (4º usuário, Tijuca,...)

O primeiro elemento identifica a pessoa. Em um CDR, o primeiro campo poderia ser o número de telefone da pessoa. O segundo campo, indicaria a região onde se tem uma torre de telefonia celular próxima da sua residência. O Apêndice I apresenta a lista dos 50 primeiros bairros sorteados. O próximo passo, segundo Isaacman et al. (2012), é escolher onde a pessoa trabalha[65].

Como o método exemplificado é o *Home and Work*, será feita uma restrição à ideia de que só serão sorteadas pessoas que trabalham e residem na mesma cidade[66] e no mesmo conjunto de bairros selecionados para análise. Isso quer dizer que elas, necessariamente, irão se deslocar da casa para o trabalho nos dias úteis e ambos estão localizados em um dos bairros isolados previamente para síntese do CDR. Na verdade, o modelo pode ficar muito mais complexo, do ponto de vista conceitual, do que o que está sendo ilustrado, contemplando a população que trabalha à noite, ou nos fins de semana, por exemplo. Mas isso deverá ser tratado em trabalhos posteriores.

Para dar continuidade à implementação da pesquisa, será sorteada uma região de trabalho para cada uma das pessoas "criadas". O método original determina que seja sorteada, para cada indivíduo criado, uma distância de trânsito, supostamente utilizada para ir ao trabalho. Uma vez determinada esta distância, é sorteado um local (região) de

[65] Isaacman et al. (2012, p.241): For each point in space, a second probability distribution Commute Distance expresses the probability of having different commute distances, conditioned on that given home location. A commute distance, d, selected from this distribution can be envisioned as describing a circle of radius d around the selected home location. Next, our method selects a work location somewhere along this circle. To do so, a third distribution, Work, gives the probability of different work locations around a circle of commute distance d from the home location. These Home and Work locations are derived from population densities that correspond to the city we wish to synthesize.

[66] A cidade do Rio de Janeiro, neste estudo, foi reduzida a estes 8 bairros, de forma que ninguém trabalhe fora dos bairros escolhidos para exemplificar o cenário prático do método WHERE nesta dissertação.

trabalho que se encontre a esta distância de sua residência. A ideia é que um indivíduo, na região da Barra da Tijuca, por exemplo, viajaria sempre uma distância média tal que, probabilisticamente, indicaria que ele trabalha nas regiões que estivessem a essa mesma distância da Barra da Tijuca. Para fixar ideias, serão usados os bairros de Botafogo, de Madureira ou de Copacabana como postos de trabalho de um determinado indivíduo. De acordo com as densidades de postos de trabalho em Botafogo, Madureira ou Copacabana, seria sorteado o seu posto de trabalho que seria, por exemplo, Copacabana.

Nesta dissertação, entretanto, optou-se por assumir que a probabilidade de alguém trabalhar em um determinado bairro entre os escolhidos, independe de onde a pessoa reside. Esta probabilidade está associada somente ao número de postos de trabalho da região sorteada.

Serão consideradas as quantidades de postos de trabalho de acordo com a Figura 9 e o Quadro 15.

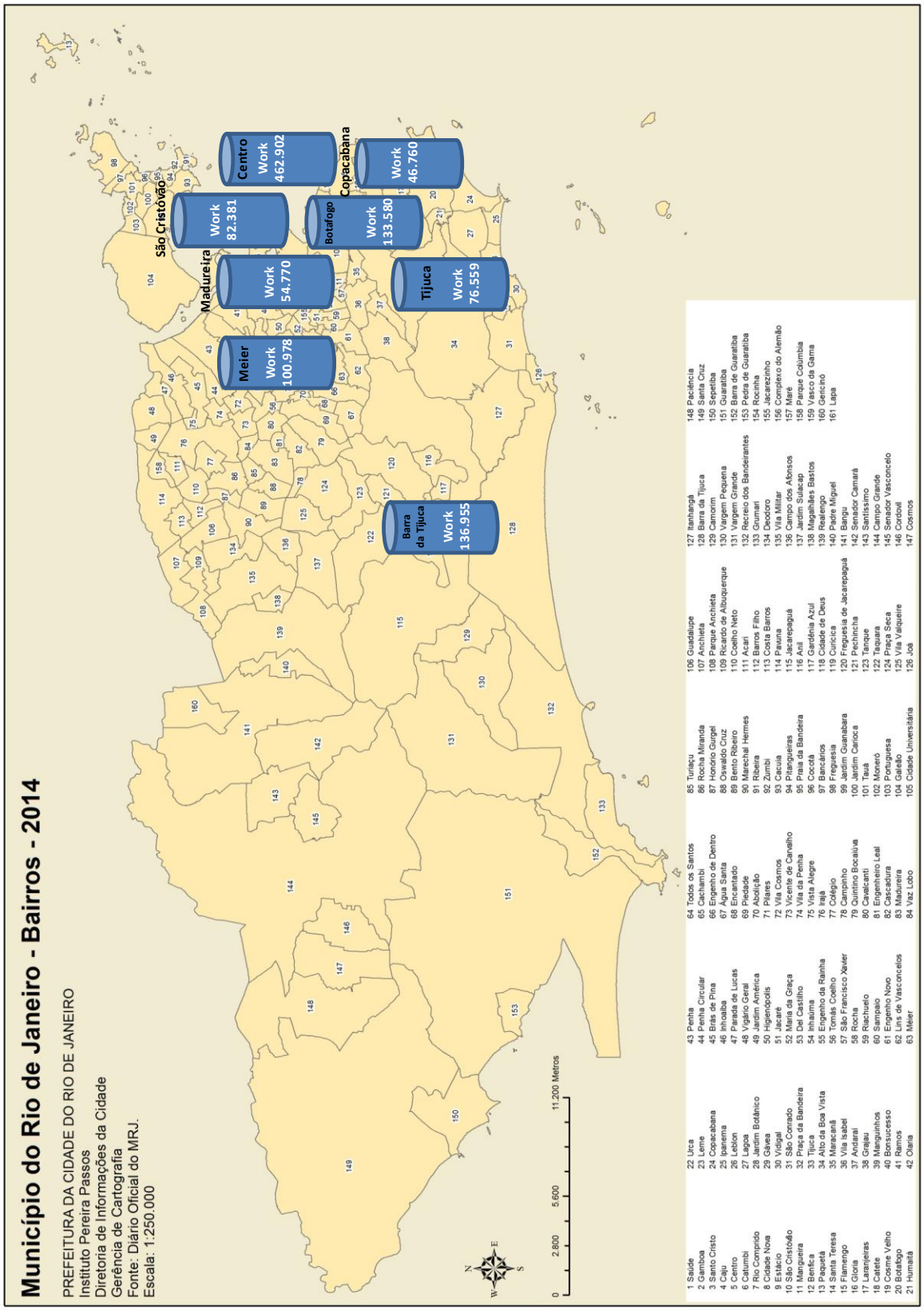


Figura 9 - Postos de trabalho por bairro do Município do Rio de Janeiro
Fonte: Instituto Municipal de Urbanismo Pereira Passos –2008

Bairros	Postos de trabalho
Barra da Tijuca	136.955
Botafogo	133.580
Centro	462.902
Copacabana	46.760
Madureira	54.770
Meier	100.978
São Cristóvão	82.381
Tijuca	76.559

Quadro 15 - Total de postos de trabalho por bairro do Município do Rio de Janeiro
Fonte: Instituto Municipal de Urbanismo Pereira Passos, 2008

O Quadro 15 apresenta o total de postos de trabalho por bairro em cada um desses lugares, totalizando 1.094.885.

Ao se sortear uma pessoa qualquer, a probabilidade dela trabalhar em um dos bairros selecionados acima será dada pela probabilidade descrita na Tabela 3.

Tabela 3 - Distribuição de probabilidade de um habitante da região isolada de residir em cada um destes bairros

Bairros	Postos de trabalho
Barra da Tijuca	13%
Botafogo	12%
Centro	42%
Copacabana	4%
Madureira	5%
Meier	9%
São Cristóvão	8%
Tijuca	7%
Total	100%

Fonte: Elaborado pelo autor

Isaacman et al. (2012) afirmam que as distâncias trafegadas precisam ser ligadas à região de moradia e citam como exemplo Manhatan e subúrbios[67]. Conforme já dito anteriormente, não se assumiu sempre uma mesma distância diária de casa ao trabalho para todas as pessoas. Ou seja, se a distância média de traslado fosse 40km, as pessoas que moram no Centro iriam viajar em média 40km também, o que seria claramente um

[67] Isaacman et al. (2012, p.243): It is not sufficient to select from a general probability distribution for commute distances, because this may unfairly bias toward commute distances for very dense areas. Intuitively, the likely commute distances for a person living in midtown Manhattan are quite different from those who live in outlying exurbs. Third, all possible locations that are the selected distance away from the chosen home location are considered as possible “work” locations, and a work location is chosen for the synthetic user. Possible work locations are weighted with probabilities given by Work, and again are conditioned on a particular home location and commute distance.

absurdo. Como não foi encontrada informação pública disponível sobre as distâncias trafegadas no trajeto casa-trabalho por bairro, procedeu-se da forma mais lógica, que é vincular a probabilidade de se trabalhar num certo bairro a quantidade de postos de trabalho que lá existirem, o que parece uma hipótese bem razoável em grandes centros, como o Rio de Janeiro.

Como este estudo não usará como referência as distâncias trafegadas de casa para o local de emprego, a probabilidade de um morador trabalhar em uma região qualquer será igual ao percentual atribuído a esta região, de acordo com a Tabela 3.

O método já utilizado para definir a residência será, então, repetido. Considerando cada bairro uma célula, foram atribuídos valores às células de acordo com a porcentagem assinalada na Tabela 3 – Distribuição de probabilidade de um habitante residir nos bairros escolhidos –, como mostra o Quadro 16.

Bairros	Postos de trabalho
Barra da Tijuca	1 a 13
Botafogo	14 a 25
Centro	26 a 67
Copacabana	68 a 71
Madureira	72 a 76
Meier	77 a 85
São Cristóvão	86 a 93
Tijuca	94 a 100

Quadro 16 - Pontuação por bairro em função da probabilidade do bairro ser local de trabalho para algum dos moradores da região isolada
Fonte: Elaborado pelo autor

Para o sorteio de uma pessoa e identificação de onde ela trabalha, entre um dos bairros referidos no Quadro 16, será utilizada a mesma função randômica do Excel que sorteou onde o indivíduo reside. A expressão exibida no Quadro 17 mostra o resultado da execução da função aplicada para sortear em que bairro a pessoa trabalha.

f(x) = TRUNCAR(ALEATÓRIO()*100)+1

Onde:

f(x) = É o resultado aleatório do bairro onde a pessoa trabalha.

Quadro 17 - Função randômica do Excel que sorteia um número de 1 a 100 num espaço equiprovável
Fonte: Elaborado pelo autor

A expressão apresentada no Quadro 17 trará valores que irão de 1 a 100. Se o número que sair for 15, saber-se-á que a pessoa trabalha em Botafogo. Caso o número sorteado seja 2, o local de trabalho será o bairro Barra da Tijuca e, assim, sucessivamente.

Desta forma, se forem sorteadas 200 pessoas, será possível saber exatamente onde cada pessoa trabalha e ter-se-á uma lista sequencial dos registros, conforme a seguir:

- (1º usuário, Barra da Tijuca, Botafogo,...)
- (2º usuário, Centro, Barra da Tijuca,...)
- (3º usuário, São Cristóvão, São Cristóvão,...)
- (4º usuário, Tijuca, Meier,...)

O Quadro 18 exibe o resultado da execução da função aplicada para sortear em que bairro uma pessoa qualquer trabalha.

Número sorteado do registro	Bairro de trabalho
15	Botafogo
2	Barra da Tijuca
90	São Cristóvão
82	Meier
...	...

Quadro 18 - Registro do sorteio dos bairros de trabalho
Fonte: Elaborado pelo autor

No estudo de questões ligadas exclusivamente à mobilidade, deve-se dar maior ênfase a esta parte inicial da produção do CDR, criando maneiras de refletir melhor os movimentos populacionais verdadeiros na criação desses indivíduos. Para este objetivo, deve-se simplificar as condições seguintes, pois o seu objetivo será somente ter um resultado final que possa ser utilizado por *softwares* já disponíveis no mercado, tornando desnecessária, portanto, a criação de instrumentos para sua análise[68].

Desta forma, para as 200 pessoas selecionadas, ter-se-á uma lista de registro do tipo:

- (1º usuário, Barra da Tijuca, Botafogo,...)
- (2º usuário, Centro, Barra da Tijuca,...)

[68] Isaacman et al. (2012, p.240): this output can plug in directly into the growing body of analysis software that uses CDRs as input.

– (3º usuário, São Cristóvão, São Cristóvão,...)
– (4º usuário, Tijuca, Meier,...)

O primeiro elemento identifica a pessoa (num CDR, poderia ser um número de telefone) e o segundo elemento revela a região onde ela mora. O terceiro elemento fornece a região onde ela trabalha – Quadro 19.

Número sequencial do registro por linha	Bairro de residência	Bairro de trabalho
1	Barra da Tijuca	Botafogo
2	Centro	Barra da Tijuca
3	São Cristóvão	São Cristóvão
4	Tijuca	Meier
...	...	...

Quadro 19 - Resultado dos sorteios de residência e de postos de trabalho
Fonte: Elaborado pelo autor

O Apêndice II apresenta a lista dos 50 primeiros bairros de trabalho sorteados. A próxima etapa irá estabelecer um comportamento médio de ligações para cada indivíduo.

Isaacman et al. (2012) utilizaram, em seu estudo, as distribuições retiradas de CDRs reais (CallTime e PerUserCallsPerDay), mas oferecem como alternativa as distribuições de outros trabalhos já publicados[69].

Como esta parte da produção do CDR interessa fundamentalmente às companhias telefônicas, os parâmetros utilizados podem ser estabelecidos de acordo com os dados que as companhias já possuem ou em função do que as companhias pretendem simular.

O valor da probabilidade pode ser único ou pode variar de bairro para bairro. Neste estudo, será considerada uma regra única, assumindo que estas probabilidades serão válidas para qualquer usuário, independentemente do bairro onde ele estiver.

Para que esta pesquisa se aproxime da realidade e utilize simulações que se aproximem do que ocorre cotidianamente, serão utilizados três perfis de ligadores[70], que

[69] Isaacman et al. (2012, p.242-243): Without access to CDR information, it is still possible to construct call distributions. One approach would be to assume that all users have similar temporal patterns, and simply use an overall per-hour call probability distribution function as the guide for when calls are most or least likely to be made. Such probabilities can be drawn from prior work including [1, 5].

[70] Ligadores são pessoas que possuem perfis diferentes e que realizam números médios de ligações telefônicas diferentes por dia.

respondem a 20%, 30% e 50% das ligações, e serão arbitrados valores de média e desvio-padrão para eles, conforme exibido no Quadro 20 e ilustrado no Gráfico 8.

Perfil	Probabilidade	Média de ligações por dia	Desvio-Padrão
X	20%	20	5
Y	30%	10	1
Z	50%	4	2

Quadro 20 - Perfil de ligadores
Fonte: Elaborado pelo autor

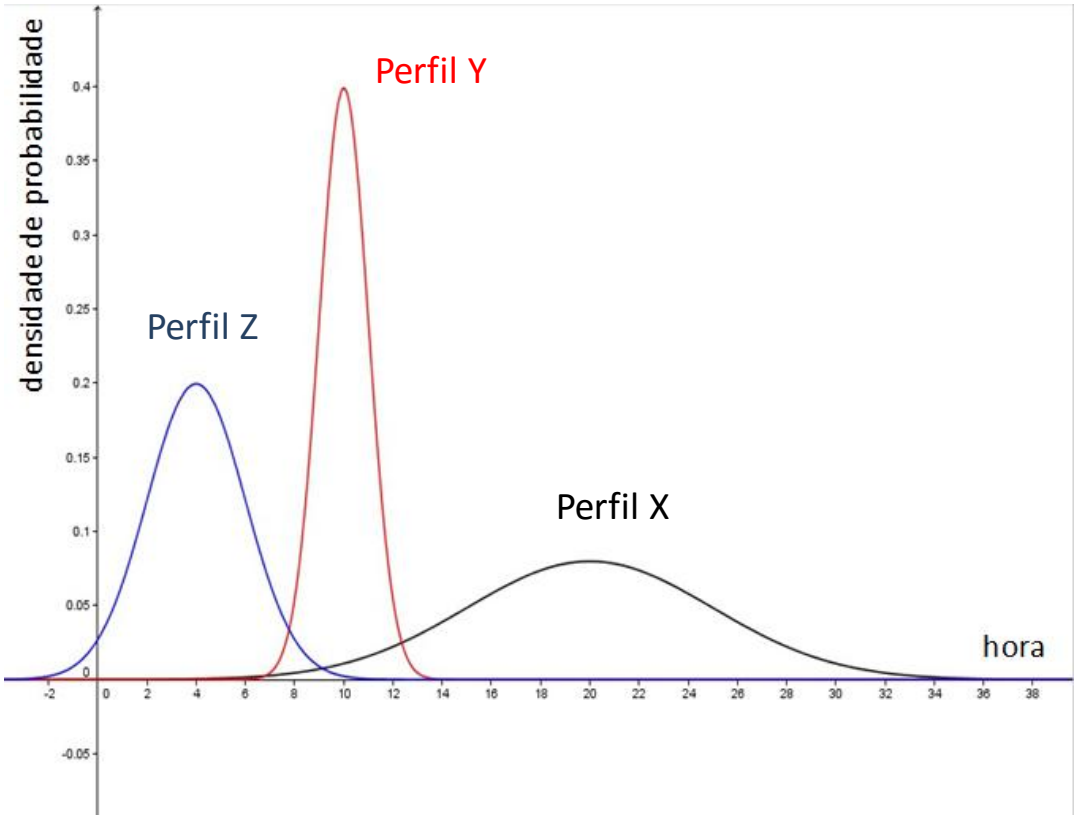


Gráfico 8 - Perfil de ligador
Fonte: Elaborado pelo autor

A expressão descrita no Quadro 21 estabelece o perfil de ligador do tipo X com probabilidade de 20% de realizações de ligações, com média de 20 ligações por dia e com desvio-padrão igual a 5.

$$f(x) = \frac{e^{\frac{-((x-20)^2)}{50}}}{5\sqrt{2\pi}}$$

Quadro 21 - Perfil de ligador do tipo X com probabilidade de 20% de realizações de ligações, com média de 20 ligações por dia e com desvio-padrão igual a 5

A função g(x), expressa na equação do Quadro 22, estabelece o perfil de ligador do tipo Y com probabilidade de 30% de realizações de ligações, com média de 10 ligações por dia e com desvio-padrão igual a 1.

$$g(x) = \frac{e^{\frac{-((x-10)^2)}{2}}}{\sqrt{2\pi}}$$

Quadro 22 - Perfil de ligador do tipo Y com probabilidade de 30% de realizações de ligações, com média de 10 ligações por dia e com desvio-padrão igual a 1

A função h(x), expressa no Quadro 23, estabelece o perfil de ligador do tipo Z com probabilidade de 50% de realizações de ligações, com média de 4 ligações por dia e com desvio padrão igual a 2.

$$h(x) = \frac{e^{\frac{-((x-4)^2)}{8}}}{2\sqrt{2\pi}}$$

Quadro 23 - perfil de ligador do tipo Z com probabilidade de 50% de realizações de ligações, com média de 4 ligações por dia e com desvio-padrão igual a 2

Em seguida, deve-se realizar o sorteio do número de ligações para cada perfil definido, conforme Quadro 24.

Perfil	Número de pessoas por perfil
X	1 a 20
Y	21 a 50
Z	51 a 100

Quadro 24 - Quantidade de pessoas por perfil
Fonte: Elaborado pelo autor

Para sortear uma pessoa e identificar a média de ligações por ela realizada diariamente, será utilizada a mesma função que foi aplicada no sorteio do bairro de residência e do bairro de trabalho e reproduzida no Quadro 25.

f(x) = TRUNCAR(ALEATÓRIO()*100)+1

Onde:

f(x) = O resultado estará relacionado ao perfil do ligador

Quadro 25 - Função randômica do Excel que sorteia um número de 1 e 100 num espaço equiprovável
Fonte: Elaborado pelo autor

Similarmente à função exposta no Quadro 17, a função apresentada no Quadro 25 trará valores que irão de 1 a 100. Os cinco primeiros números sorteados foram: 68, 71, 99, 2 e 37.

No exemplo acima, os números sorteados pertencem aos perfis discriminados no Quadro 26.

Número sorteado	Perfil de ligador
68	Z
71	Z
99	Z
2	X
37	Y

Quadro 26 – Números sorteados por perfil selecionado
Fonte: Elaborado pelo autor

Em função do sorteio realizado, a lista de montagem de CDR sintético (hipotético) passa a ter a seguinte formação:

- (1º usuário, Barra da Tijuca, Botafogo, Z,..)
- (2º usuário, Centro, Barra da Tijuca, Z,...)
- (3º usuário, São Cristóvão, São Cristóvão, Z,...)
- (4º usuário, Tijuca, Meier, X,...)

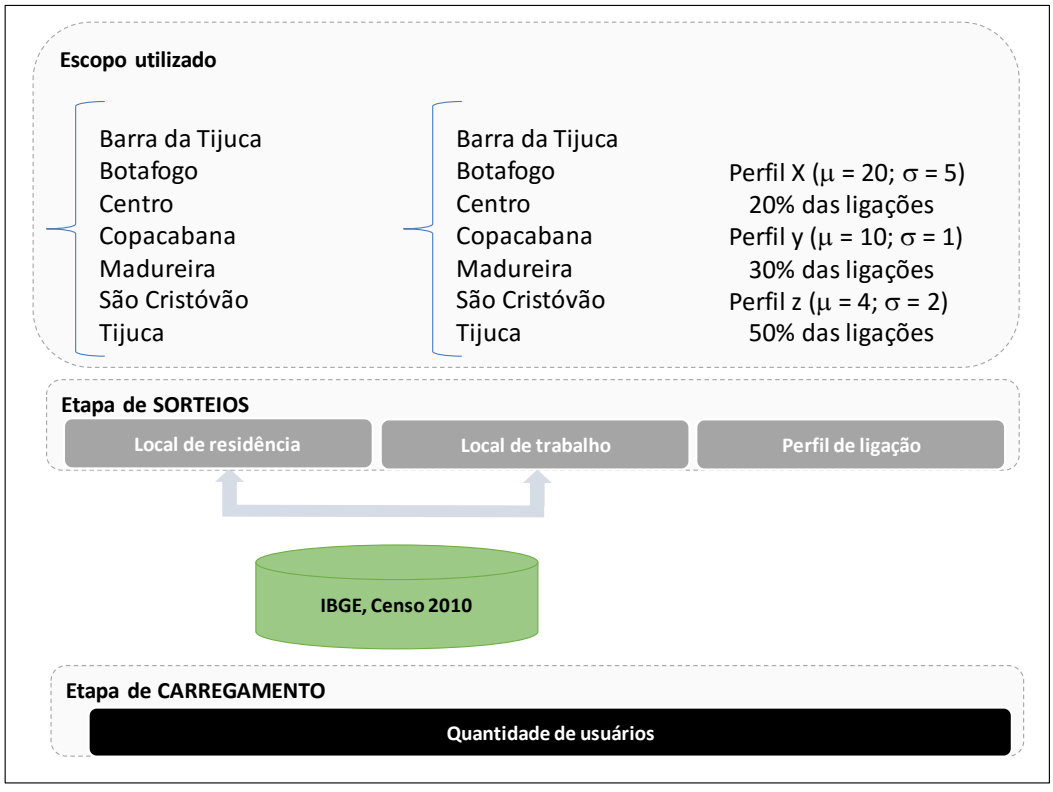
Na relação anterior, o primeiro elemento identifica a pessoa (num CDR, poderia ser um número de telefone), o segundo elemento revela a região onde ela mora, o

terceiro elemento é a região onde ela trabalha e o quarto, o perfil de ligador associado a essa pessoa (Quadro 27).

Número sequencial do registro por linha	Bairro de residência	Bairro de trabalho	Perfil do ligador
1	Barra da Tijuca	Botafogo	Z
2	Centro	Barra da Tijuca	Z
3	São Cristóvão	São Cristóvão	Z
4	Tijuca	Meier	X
...	...	...	Y

Quadro 27 – Probabilidade de perfil de ligadores
Fonte: Elaborado pelo autor

Toda esta parte refere-se ao primeiro algoritmo, que é o *Create* – etapa [1][71] – ver Quadro 28.



Quadro 28- Fluxo de produção de CDR sintético
Fonte: Elaborado pelo autor

[71] As etapas [5] e [6] do algoritmo *Create* foram aglutinadas em um único passo, que termina por fornecer, da mesma forma, a média e o desvio-padrão da quantidade de ligações por dia do referido usuário. Também foram condensados em um único passo as etapas [3] e [4].

Em seguida, foi sorteada uma sequência de cinco números com as quantidades de ligações realizadas pelo primeiro usuário em cada um dos cinco dias úteis da semana. Partiu-se da premissa de que as quantidades de ligações diárias de cada usuário se encaixam em distribuições normais, com média e desvio-padrão do respectivo perfil.

O sorteio, de acordo com a média e desvio-padrão fornecidos, pode ser feito pela função do Excel exibida no Quadro 29.

f(x) = ARRED(INV.NORM(ALEATÓRIO();**4**;**2**);0)

Onde:

*Os números **4** e **2**, respectivamente, são a média e o desvio padrão da distribuição, associada ao perfil considerado.*

f(x) = nº médio de ligações do ligador.

Quadro 29 - Sorteia um número de ligações associado a determinado perfil (no caso, o perfil Z)
Fonte: Elaborado pelo autor

Utilizando a função acima com, por exemplo, parâmetros 4 (média) e 2 (desvio-padrão), são gerados cinco valores que correspondem às quantidades de chamadas em cada um dos cinco dias da semana de trabalho[72, 73] considerados.

A função foi escolhida para fornecer números que se ajustam à distribuição gaussiana, com a média e o desvio-padrão indicados. Por exemplo: para o perfil X, com média 20 e desvio-padrão 5, sorteando 100 números, os dados sorteados compõem o histograma representado pelo Gráfico 9, revelando a forma de sino típica de uma distribuição normal.

[72] Cinco dias úteis da semana de trabalho são: segunda, terça, quarta, quinta e sexta-feira.
[73] Isaacman et al. (2012, p.243): Finally, after having selected a user's home and work, the synthetic user is assigned a calling pattern (i.e., mean μ and standard deviation s calls per day) according to the distributions from CallTime and PerUserCallsPerDay.

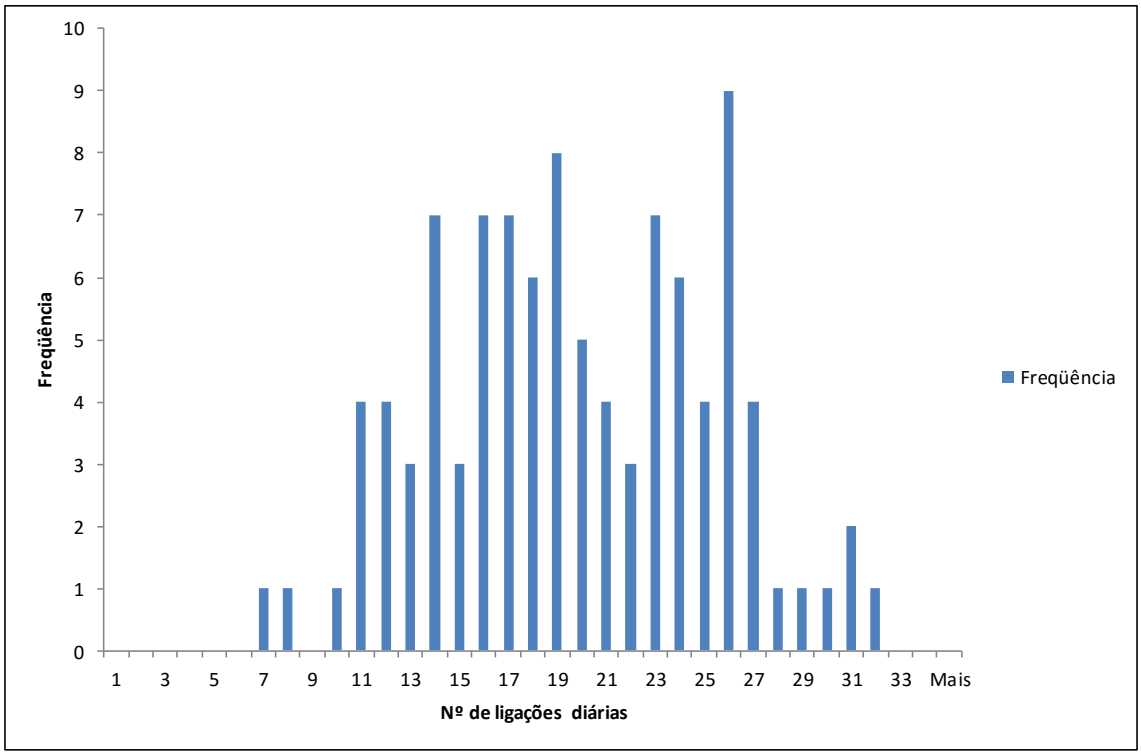


Gráfico 9 - Distribuição nº de ligações diárias para o perfil de ligador X
Fonte: Elaborado pelo autor

Prosseguindo o exemplo, para o primeiro registro da lista (1º usuário, Barra da Tijuca, Botafogo, Z ...), foram geradas quantidades de chamadas para ele para os cinco dias úteis da semana. A quantidade de dias é imaterial. Aquele que estiver interessado em utilizar o método para simular chamadas telefônicas poderá alterar este ponto do método para que o perfil de chamadas, por exemplo, dependa do dia da semana, uma vez que é bastante intuitivo que o padrão de chamadas (periodicidade, quantidade e duração) em um fim de semana não seja idêntico ao padrão durante a semana. Tem-se, assim, o resultado mostrado no Quadro 30.

2ª feira	3ª feira	4ª feira	5ª feira	6ª feira	Média de ligações	Desvio-padrão
5	4	5	4	2	4	2

Quadro 30 – Número de ligações de ligações dos usuários por dias da semana
Fonte: Elaborado pelo autor

Para os dois próximos sorteios do perfil Z, foram obtidos os seguintes resultados (Quadro 31):

2ª feira	3ª feira	4ª feira	5ª feira	6ª feira	Média de ligações	Desvio-padrão
2	0	6	2	7	4	2
3	3	4	2	1	4	2

Quadro 31 - Número de ligações dos usuários de perfil Z
Fonte: Elaborado pelo autor

Para o quarto sorteio no perfil X, foram usados como parâmetros 20 (média de ligações) e 5 (desvio-padrão), gerando uma amostra de cinco valores, correspondente aos cinco dias úteis da semana de trabalho[74], com o resultado exibido no Quadro 32.

2ª feira	3ª feira	4ª feira	5ª feira	6ª feira	Média de ligações	Desvio-padrão
16	20	23	24	19	20	5

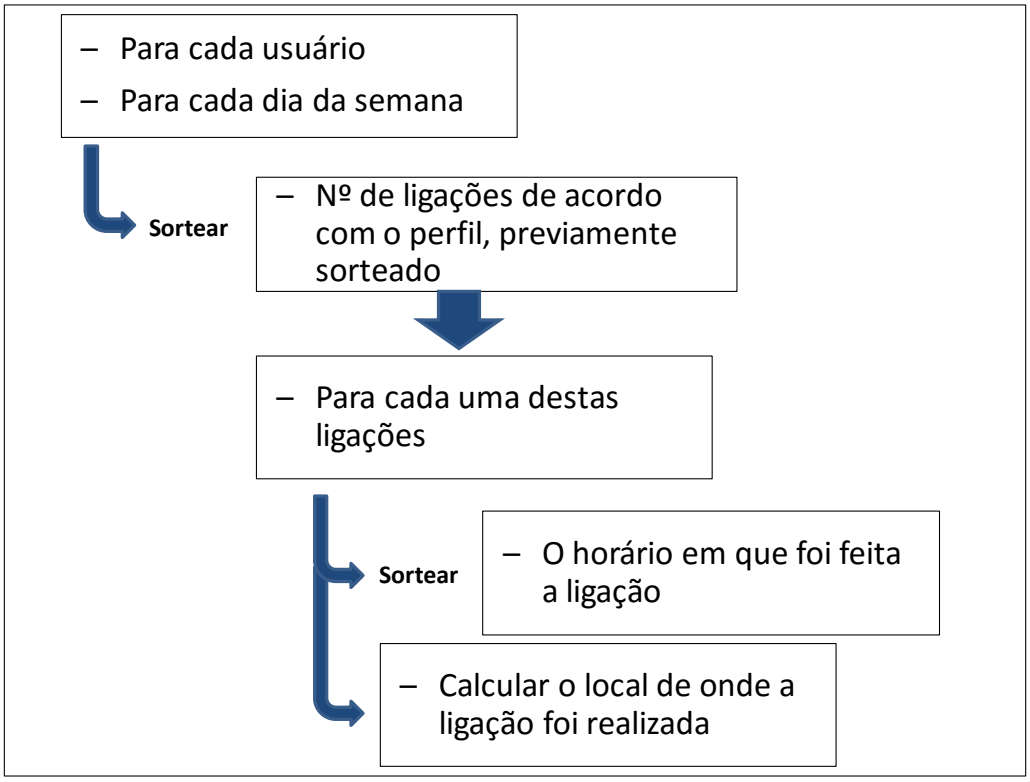
Quadro 32 - Número de ligações dos usuários no perfil X
Fonte: Elaborado pelo autor

O primeiro registro de ligação da lista de montagem de CDR sintético ficaria da seguinte forma:

- (1º usuário, Barra da Tijuca, Botafogo, Z, (5,4,5,4,2),...)
- (2º usuário, Centro, Barra da Tijuca, Z, (2,0,6,2,7),...)
- (3º usuário, São Cristóvão, São Cristóvão, Z, (3,3,4,2,1),...)
- (4º usuário, Tijuca, Meier, X, (16,20,23,24,19),...)

A construção dessa lista já é o primeiro passo do algoritmo *Move* (Quadro 33).

[74] Cinco dias úteis da semana de trabalho são: segunda, terça, quarta, quinta e sexta-feira.



Quadro 33 - Fluxo de produção de CDR sintético – Algoritmo *Move*
Fonte: Elaborado pelo autor

Nesse algoritmo, percorre-se a lista de usuários e, para cada elemento da lista, percorrem-se os dias para os quais se deseja gerar registros.

A questão agora é: Como sortear os horários em que cada uma destas ligações foi realizada?

A recomendação do artigo de Isaacman et al. (2012)[75] é que, na ausência de dados reais, sejam utilizadas distribuições de outros trabalhos, como o de Queijo e Almeida (1998).

O trabalho de Queijo e Almeida (1998) infere, por experiência feita em Lisboa, que uma boa distribuição para se utilizar na previsão de horários de ligações na maior parte das zonas classificadas (Quadro 34) é uma dupla gaussiana.

[75] Isaacman et al. (2012, p. 243): To summarize, the sources of temporal input data on call patterns can either be published statistics [1, 5] or proprietary CDR data. Because these call patterns have been seen multiple times in multiple contexts, we assume that such a calling pattern is general enough to hold regardless of the spatial data to which it is applied.

Centro Urbano (C)
Centro Urbano com Estradas (CE)
Residencial (R)
Residencial com Estradas (RE)
Suburbana (S)
Suburbana com Estradas (SE)

Quadro 34 - Locais geográficos
Fonte: Queijo e Almeida, 1998

A dupla gaussiana é modelada pela função mostrada no Quadro 35.

$$ap_{gauss\,(t_{desv})} = \begin{cases} p_1 . e^{-\frac{(t_{desv} - h1_{desv})^2}{2d_1^2}} & t_{desv \leq h_{alm\,desv}} \\ p_2 . e^{-\frac{(t_{desv} - h_{2desv})^2}{2d_2^2}} & t_{desv > h_{alm\,desv}} \end{cases}$$

Quadro 35 - Função dupla gaussiana
Fonte: Queijo e Almeida, 1998

Nesta equação, os parâmetros utilizados são:

P_1 – é a amplitude da primeira gaussiana;

$H_{1\,desv}$ – é a hora de pico da manhã desviada;

D_1 – é o desvio-padrão da primeira gaussiana;

$H_{alm\,desv}$ – é a hora de almoço desviada;

P_2 – é a amplitude da segunda gaussiana;

$H_{2\,desv}$ – é a hora de pico da tarde desviada;

D_2 – é o desvio-padrão da segunda gaussiana.

Os valores temporais foram desviados, de acordo com a função exibida no Quadro 36 para tornar mais simples a fórmula algébrica da função.

Os valores temporais desviados obtêm-se a partir de:

$$t_{desv}(t) = \begin{cases} t - 5 \; se \; 5 \leq t < 24. \\ t - 5 + 24 \; se \; 0 \leq t < 5 \end{cases}$$

Quadro 36 - Valores temporais desviados
Fonte: Queijo e Almeida, 1998

Onde (t) é medido em horas.

Abaixo, o Gráfico 10 relaciona o valor de t_{desv} (eixo Y) com o valor real de t (eixo X).

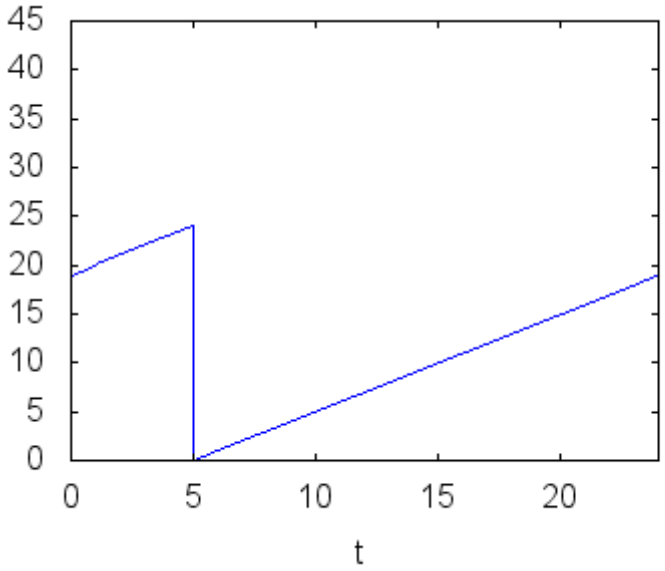


Gráfico 10 - Valor de tempo desviado em relação ao tempo real
Fonte: Elaborado pelo autor

A Tabela 4 mostra alguns exemplos de valores de parâmetros utilizados no trabalho de Queijo e Almeida (1998). Vale ressaltar que os tempos são reais; não estão desviados.

Tabela 4 - Parâmetros dos modelos normalizados de duas gaussianas

Classe	h_1 [h]	h_{alm} [h]	h_2 [h]	p_1	p_2	d_1 [h]	d_2 [h]
C	11,7	13,1	16,7	0,91	0,94	2,11	4,32
CE	11,6	13	16,4	0,92	0,96	2,17	4,04
R	11,5	13,3	16,9	0,93	0,96	2,18	4,34
RE	11,3	13	17	0,91	0,96	2,13	4,16
S	11,3	13,3	16,9	0,9	0,88	2,16	4,07
SE	11,1	13,2	17,1	0,93	0,94	2,08	3,94

Fonte: Queijo e Almeida, 1998

Utilizando os parâmetros da primeira linha da Tabela 4, foi plotado, como exemplo, o Gráfico 11 da dupla gaussiana, utilizando o *software* Maxima[76].

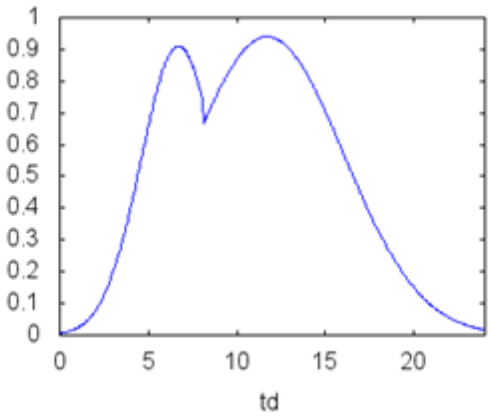


Gráfico 11 - Dupla gaussiana
Fonte: Queijo e Almeida, 1998[77]

A ideia é que há dois momentos do dia que concentram as probabilidades de que uma ligação seja feita, que são os picos das gaussianas que, no exemplo deste estudo são 11,7h (11h42min) e 16,7h (16h42min). No Gráfico 11, os tempos aparecem desviados de 5 horas, conforme a Tabela 4.

Partiu-se da ideia de que as pessoas estão no seu local de trabalho das 9h às 18h. Este período poderia, também, ser sorteado em função do perfil local. Mas optou-se pela hipótese simplificadora de que todos os habitantes trabalham e, adotando os critérios de Queijo e Almeida (1998), o perfil de ligação utilizado é o de pessoas que trabalham diurnamente em horário habitual, sendo este o perfil empregado para todos os usuários.

[76] O Maxima é um sistema de computação algébrica, desenvolvido na década de 60 no Instituto de tecnologia de Massachusetts; é de caráter livre e desenvolvido com esforço da comunidade da Internet.
[77] Queijo e Almeida (1998, p.26): Classe C da tabela de parâmetros dos modelos normalizados de duas gaussianas.

A ideia agora é bem simples: para cada uma das pessoas, para cada um dos dias, para cada uma das ligações que esta pessoa fizer naquele dia, será sorteado um horário específico para a sua ligação. Se o horário sorteado estiver dentro do horário de trabalho, pode-se afirmar que a pessoa o fez do seu local de trabalho previamente sorteado. Caso contrário, pode-se dizer que ela o fez do seu local de moradia, também previamente sorteado. É claro que este modelo poderia ser mais sofisticado, interpondo horários de trânsito (antes e depois do expediente), tempos estes em que a ligação seria feita de algum lugar entre a moradia e o trabalho. Um modelo como este é bem viável (na cidade do Rio de Janeiro, por exemplo), uma vez que há dados sobre o tempo médio de traslado de casa para o trabalho – e vice-versa – na maior parte dos bairros do Rio de Janeiro, embora exija algum cuidado[78].

O problema técnico a ser enfrentado é que este sorteio precisa ser feito de acordo com a função de probabilidade associada à dupla gaussiana mostrada no Gráfico 11. Essa dificuldade já foi contornada em outras partes deste trabalho, utilizando-se funções já prontas do Excel:

1) Quando os locais de trabalho e casa foram sorteados, o problema foi modelado de forma a ajustá-lo a um espaço equiprovável e utilizou-se uma função do Excel que fornecia os resultados de uma variável aleatória dentro do escopo daquele espaço.
2) Quando as quantidades de ligações em cada dia da semana foram sorteadas, utilizou-se uma função do Excel que produzia os resultados de uma variável aleatória adaptados ao escopo de uma gaussiana.

Para superar esta dificuldade, é necessário, de acordo com Olver e Townsend (2006), primeiramente, criar uma função acumulada de probabilidade, que, no caso da dupla gaussiana apresentada no Gráfico 11, seria o valor da área entre o gráfico e o eixo horizontal de zero a um tempo arbitrário x, normalizando-a para que a área total ao longo de um dia seja 1 (uma unidade, ou 100%). A área não normalizada (w(x)) é obtida a partir de uma integração da dupla gaussiana, de zero até um tempo arbitrário x.

No Maxima (Apêndice III), foi utilizado o comando para gerar w(x): ***w(x)=quad_qags(ap_gauss(td), td, 0, x).***

[78] A fim de evitar, por exemplo, o caso de uma pessoa que more e trabalhe no bairro da Barra da Tijuca e gaste o mesmo tempo para se deslocar para o seu local de trabalho que uma pessoa que more num bairro e trabalhe em outro muito distante e de difícil acesso ou, pior que isso, o caso de uma pessoa que para vencer a distância entre casa e trabalho no tempo escolhido tivesse que se locomover, por exemplo, a 180km/h.

A função de probabilidade acumulada é achada dividindo-se w(x) pela área total entre a gaussiana e o eixo horizontal. Esta função de probabilidade acumulada, que na folha de trabalho do Maxima foi obtida com o comando ***prob_ac2(x):=(w(x))[1]/área*** , gerará a função P(x) – Gráfico 12. Detalhes podem ser encontrados no Anexo V - Folha de trabalho gerado pelo Maxima.

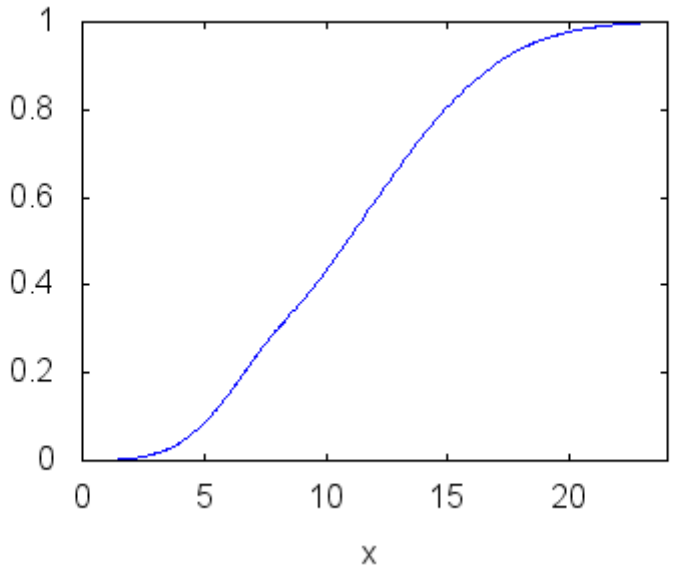


Gráfico 12 – Função P(x), em x em horas, de probabilidade acumulada da dupla gaussiana utilizada
Fonte: Elaborado pelo autor

Em seguida, ainda de acordo com Olver e Townsend (2006), foram sorteados valores aleatórios dentro de um espaço equiprovável (neste estudo, foi empregada, por comodidade, uma função nativa do Maxima, mas poder-se-ia usar qualquer outra, inclusive a já utilizada do Excel), x_1, x_2, x_3, ..., x_n , e as equações $P(y_i)=x_i$ foram resolvidas.

Os valores de y_i serão exatamente os correspondentes das variáveis x_i dentro do espaço adequado à distribuição de probabilidade da dupla gaussiana utilizada. Estes valores de y_i correspondem aos horários desviados[79] das ligações sorteadas.

A solução de cada equação foi obtida com o comando abaixo, do Maxima. Nesta expressão, que na folha de trabalho encontra-se dentro de um *loop*, ***random(1.0)*** assumirá um valor diferente para cada iteração do *loop*, representando os valores de x_i, que estão num espaço amostral equiprovável. O resultado de cada iteração irá fornecer o

[79] Horários Desviados são a imagem da função t_{desv}. São necessários para facilitar a abordagem algébrica da dupla gaussiana.

respectivo valor para y_i, no espaço de probabilidades desejado (da dupla gaussiana), como foi descrito no Capítulo 4: ***find_root (random(1.0)=prob_ac2(y_i),y_i,0,24)***

Como conjunto de parâmetros, foi utilizada a primeira linha da tabela de parâmetros dos modelos normalizados de duas gaussianas (Tabela 4), mas o conjunto específico não é tão importante assim. Num modelo mais complexo, o mais importante é que sejam bem determinadas as horas de pico, a hora de almoço, a relação entre as amplitudes das gaussianas (se o usuário liga muito mais à tarde, por exemplo, p2 seria maior que p1) e os desvios (se as ligações à tarde são mais concentradas em torno do horário de pico do que as ligações da manhã, por exemplo, d2 seria menor que d1).

Os valores de d2 e d1 precisam ser coerentes com os valores atribuídos aos horários de pico, de acordo com a ideia acadêmica de desvio-padrão conhecida, mas as amplitudes em si podem ser imateriais, visto que o processo de sorteio utilizado normalizará os dados.

Ou seja, poder-se-ia utilizar p1=1; p2=1,5; d1 = 2h e d2 = 0,5h. Isso daria uma distribuição em que o ligador possui um pico de ligações à tarde 50% maior do que de manhã e muito mais concentradas em torno do horário da tarde. Veja o Gráfico 13 com horários desviados.

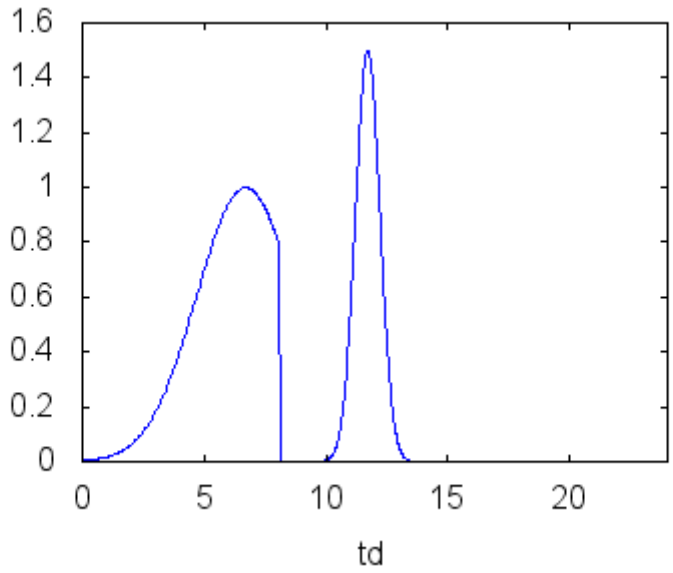


Gráfico 13 – Dupla Gaussiana (horários desviados, em horas, de acordo com a função t_{desv} , definida no Quadro 36)
Fonte: Elaborado pelo autor

Esta flexibilidade permite à pessoa que modela a população conseguir conferir uma diversidade mais realística na escolha dos seus usuários.

Os primeiros registros de ligações da lista de montagem de CDR sintético são os seguintes:

- (1º usuário, Barra da Tijuca, Botafogo, Z, (5,4,5,4,2),...)
- (2º usuário, Centro, Barra da Tijuca, Z, (2,0,6,2,7),...)
- (3º usuário, São Cristóvão, São Cristóvão, Z, (3,3,4,2,1),...)
- (4º usuário, Tijuca, Meier, X, (16,20,23,24,19),...)

O Quadro 37 apresenta os campos do cabeçalho do registro de ligações telefônica do CDR sintético.

ID	Local de origem	Horário da ligação	Duração da chamada	Horário de término	Data do Início
Identificação do usuário	Identificação do local da residência ou do trabalho da origem da chamada	00:00:00	00:00:00	00:00:00	xx/xx/xxxx

Quadro 37 - Cabeçalho do CDR Sintético
Fonte: Elaborado pelo autor

A identificação do local da residência ou do trabalho no CDR original é a localização da antena (latitude e longitude) mais próxima que processou a chamada telefônica. No exemplo desta dissertação, para facilitar a leitura, utiliza-se um código mnemônico com dois caracteres para identificar o bairro (Botafogo = BF, Barra da Tijuca = BT, Centro = CE, São Cristóvão = SC, Madureira = MD, Méier = ME, Tijuca= TJ), seguidos das letras “C” (casa) ou “T” (trabalho) e de um número capaz de identificar o usuário (1,2,3 etc.). Assim, tem-se, por exemplo, que BFT1 quer dizer “local de Trabalho do usuário 1, que trabalha em Botafogo).

Para aproximar ainda mais o CDR sintético (hipotético) do CDR real, no campo identificação da localização da origem da chamada, pode-se utilizar as localizações de torres de celulares (BTS 2G, Node-B 3G e eNode-B 4G) disponíveis publicamente em *sites* como o OpenCellid (http://opencellid.org).

Como não foi tratado no método original, definiu-se a duração das chamadas como sendo de três minutos. Esse valor também pode ser variável de uma ligação para outra e mais um item de sorteio, caso seja conveniente para implementação futura de novos trabalhos.

A primeira coluna representa a identificação de um mesmo usuário pelas várias linhas do CDR, que pode ser um número de telefone da origem da chamada (ligador).

No escopo deste trabalho, também para facilitar a leitura, são utilizados, simplesmente, os primeiros números naturais (001,002,003 etc.).

Dando sequência à montagem do CDR sintético dos primeiros 50[80] registros de chamadas telefônicas com os horários em que cada uma das ligações foi realizada, de acordo com as premissas[81] estabelecidas, tem-se, então:

001;BTC1;18:10:29;00:03:00;18:13:29;08/06/2015
001;BFT1;13:05:22;00:03:00;13:08:22;08/06/2015
001;BTC1;23:57:14;00:03:00;00:00:14;08/06/2015
001;BFT1;14:37:34;00:03:00;14:40:34;08/06/2015
001;BFT1;16:36:15;00:03:00;16:39:15;08/06/2015
001;BFT1;10:40:19;00:03:00;10:43:19;09/06/2015
001;BFT1;12:04:56;00:03:00;12:07:56;09/06/2015
001;BFT1;09:25:21;00:03:00;09:28:21;09/06/2015
001;BTC1;20:25:17;00:03:00;20:28:17;09/06/2015
001;BFT1;15:59:01;00:03:00;16:02:01;10/06/2015
001;BTC1;19:00:05;00:03:00;19:03:05;10/06/2015
001;BFT1;12:35:02;00:03:00;12:38:02;10/06/2015
001;BTC1;00:38:18;00:03:00;00:41:18;10/06/2015
001;BTC1;20:54:12;00:03:00;20:57:12;10/06/2015
001;BTC1;21:12:35;00:03:00;21:15:35;11/06/2015
001;BFT1;12:18:27;00:03:00;12:21:27;11/06/2015
001;BTC1;22:24:25;00:03:00;22:27:25;11/06/2015
001;BTC1;00:09:55;00:03:00;00:12:55;11/06/2015
001;BFT1;12:01:46;00:03:00;12:04:46;12/06/2015
001;BTC1;19:01:15;00:03:00;19:04:15;12/06/2015
002;CEC2;06:12:00;00:03:00;06:15:00;08/06/2015
002;BTT2;17:51:24;00:03:00;17:54:24;08/06/2015
002;BTT2;10:25:54;00:03:00;10:28:54;10/06/2015
002;BTT2;15:04:19;00:03:00;15:07:19;10/06/2015
002;BTT2;15:02:35;00:03:00;15:05:35;10/06/2015
002;CEC2;19:22:56;00:03:00;19:25:56;10/06/2015
002;CEC2;18:37:25;00:03:00;18:40:25;10/06/2015
002;CEC2;23:09:22;00:03:00;23:12:22;10/06/2015
002;BTT2;17:06:20;00:03:00;17:09:20;11/06/2015
002;BTT2;11:36:12;00:03:00;11:39:12;11/06/2015
002;CEC2;18:12:49;00:03:00;18:15:49;12/06/2015
002;BTT2;11:16:43;00:03:00;11:19:43;12/06/2015
002;BTT2;12:23:14;00:03:00;12:26:14;12/06/2015
002;BTT2;14:49:34;00:03:00;14:52:34;12/06/2015
002;CEC2;22:33:21;00:03:00;22:36:21;12/06/2015
002;CEC2;20:42:30;00:03:00;20:45:30;12/06/2015

[80] Os 50 registros referem-se ao total de ligações que foram realizadas pelos três primeiros usuários do sorteio ao longo de uma semana útil.

[81] (i) Horário de trabalho durante a semana será das 9h00 às 18h00, (ii) Duração da chamada telefônica igual a 3 minutos, (iii) Posição Latitude e Longitude será igual para todos os lugares no mesmo bairro de residência (moradia).

002;BTT2;10:04:07;00:03:00;10:07:07;12/06/2015
003;SCT3;16:36:27;00:03:00;16:39:27;08/06/2015
003;SCT3;16:15:09;00:03:00;16:18:09;08/06/2015
003;SCT3;15:29:21;00:03:00;15:32:21;08/06/2015
003;SCT3;15:32:38;00:03:00;15:35:38;09/06/2015
003;SCT3;16:06:11;00:03:00;16:09:11;09/06/2015
003;SCC3;20:43:57;00:03:00;20:46:57;09/06/2015
003;SCT3;15:18:02;00:03:00;15:21:02;10/06/2015
003;SCT3;13:07:53;00:03:00;13:10:53;10/06/2015
003;SCT3;09:18:29;00:03:00;09:21:29;10/06/2015
003;SCT3;15:16:32;00:03:00;15:19:32;10/06/2015
003;SCT3;13:32:00;00:03:00;13:35:00;11/06/2015
003;SCT3;13:17:32;00:03:00;13:20:32;11/06/2015
003;SCT3;12:51:25;00:03:00;12:54:25;12/06/2015
...

O Quadro 38 apresenta as ligações agrupadas pelos horários de picos, das duas gaussianas.

id usuário	id local origem	horário inicial da chamada	duração da chamada	horário de término da chamada	data da ligação
1	BTC1	00:09:55	00:03:00	00:12:55	11/06/2015
1	BTC1	00:38:18	00:03:00	00:41:18	10/06/2015
2	CEC2	06:12:00	00:03:00	06:15:00	08/06/2015
3	SCT3	09:18:29	00:03:00	09:21:29	10/06/2015
1	BFT1	09:25:21	00:03:00	09:28:21	09/06/2015
2	BTT2	10:04:07	00:03:00	10:07:07	12/06/2015
2	BTT2	10:25:54	00:03:00	10:28:54	10/06/2015
1	BFT1	10:40:19	00:03:00	10:43:19	09/06/2015
2	BTT2	11:16:43	00:03:00	11:19:43	12/06/2015
2	BTT2	11:36:12	00:03:00	11:39:12	11/06/2015
1	BFT1	12:01:46	00:03:00	12:04:46	12/06/2015
1	BFT1	12:04:56	00:03:00	12:07:56	09/06/2015
1	BFT1	12:18:27	00:03:00	12:21:27	11/06/2015
2	BTT2	12:23:14	00:03:00	12:26:14	12/06/2015
1	BFT1	12:35:02	00:03:00	12:38:02	10/06/2015
3	SCT3	12:51:25	00:03:00	12:54:25	12/06/2015
1	BFT1	13:05:22	00:03:00	13:08:22	08/06/2015
3	SCT3	13:07:53	00:03:00	13:10:53	10/06/2015
3	SCT3	13:17:32	00:03:00	13:20:32	11/06/2015
3	SCT3	13:32:00	00:03:00	13:35:00	11/06/2015
1	BFT1	14:37:34	00:03:00	14:40:34	08/06/2015
2	BTT2	14:49:34	00:03:00	14:52:34	12/06/2015

(Quadro 38 – continuação)

id usuário	id local origem	horário inicial da chamada	duração da chamada	horário de término da chamada	data da ligação
2	BTT2	15:02:35	00:03:00	15:05:35	10/06/2015
2	BTT2	15:04:19	00:03:00	15:07:19	10/06/2015
3	SCT3	15:16:32	00:03:00	15:19:32	10/06/2015
3	SCT3	15:18:02	00:03:00	15:21:02	10/06/2015
3	SCT3	15:29:21	00:03:00	15:32:21	08/06/2015
3	SCT3	15:32:38	00:03:00	15:35:38	09/06/2015
1	BFT1	15:59:01	00:03:00	16:02:01	10/06/2015
3	SCT3	16:06:11	00:03:00	16:09:11	09/06/2015
3	SCT3	16:15:09	00:03:00	16:18:09	08/06/2015
1	BFT1	16:36:15	00:03:00	16:39:15	08/06/2015
3	SCT3	16:36:27	00:03:00	16:39:27	08/06/2015
2	BTT2	17:06:20	00:03:00	17:09:20	11/06/2015
2	BTT2	17:51:24	00:03:00	17:54:24	08/06/2015
1	BTC1	18:10:29	00:03:00	18:13:29	08/06/2015
2	CEC2	18:12:49	00:03:00	18:15:49	12/06/2015
2	CEC2	18:37:25	00:03:00	18:40:25	10/06/2015
1	BTC1	19:00:05	00:03:00	19:03:05	10/06/2015
1	BTC1	19:01:15	00:03:00	19:04:15	12/06/2015
2	CEC2	19:22:56	00:03:00	19:25:56	10/06/2015
1	BTC1	20:25:17	00:03:00	20:28:17	09/06/2015
2	CEC2	20:42:30	00:03:00	20:45:30	12/06/2015
3	SCC3	20:43:57	00:03:00	20:46:57	09/06/2015
1	BTC1	20:54:12	00:03:00	20:57:12	10/06/2015
1	BTC1	21:12:35	00:03:00	21:15:35	11/06/2015
1	BTC1	22:24:25	00:03:00	22:27:25	11/06/2015
2	CEC2	22:33:21	00:03:00	22:36:21	12/06/2015
2	CEC2	23:09:22	00:03:00	23:12:22	10/06/2015
1	BTC1	23:57:14	00:03:00	00:00:14	08/06/2015

Quadro 38 - Registro de CDR ordem cronológica das chamadas
Fonte: Elaborado pelo autor

Observa-se que as linhas marcadas em verde representam ligações que foram realizadas em torno do primeiro horário de pico (+/- 2,5 horas), que ocorreu às 11h 42min. Esse horário representa a primeira crista da dupla gaussiana. As linhas sombreadas em amarelo representam ligações que foram realizadas em torno do segundo horário de pico (+/- 2,5 horas), que ocorreu às 16h 42min. Esse segundo

horário representa a segunda crista da dupla gaussiana. É importante ressaltar que os registros de CDR são registrados cronologicamente. A tabela completa dos horários sorteados está descrita no Apêndice IV.

O Gráfico 14 apresenta o total de ligações realizadas por bairro de origem das ligações. Pode-se observar que a maior parte das ligações foram realizadas no horário de trabalho e dos bairros que possuem maior número de postos de trabalho.

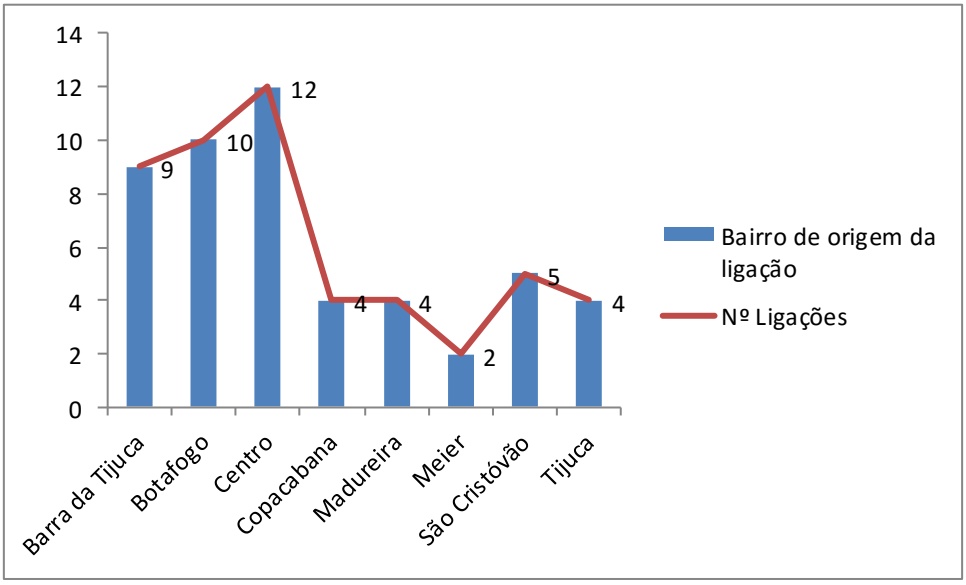


Gráfico 14 - Quantidades de ligações por bairros
Fonte: Elaborado pelo autor

O Gráfico 15 indica que a maior parte das ligações foi, de fato, realizada no horário de trabalho.

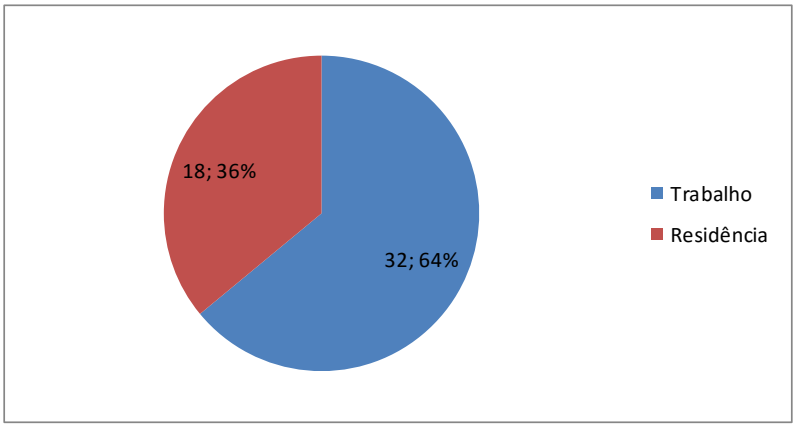


Gráfico 15 - Total de ligações no horário de trabalho e na residência
Fonte: Elaborado pelo autor

6.3 ENRIQUECENDO A GERAÇÃO DE CDR SINTÉTICO ADAPTADA

Nesta seção, serão incluídos outros campos no CDR sintético para demonstrar a flexibilidade do método de geração de CDR sintético. Esse enriquecimento aproxima ainda mais a similaridade entre os CDRs gerados automaticamente pelas centrais ou elementos de redes das operadoras aos CDRs sintéticos gerados através do método apresentado nesta dissertação.

O primeiro passo foi adicionar ao CDR sintético os campos de identificação da localização geográfica da torre de origem das chamadas (latitude e longitude), como explicado anteriormente. Essas informações podem ser obtidas publicamente pelo acesso ao *site* OpenCellid (www.opencellid.org).

No passo seguinte, identificou-se o local físico em que a ligação foi realizada, por meio da informação de latitude e longitude, ou seja, o local no bairro em que a chamada foi originada.

Reproduzindo a mesma sequência da montagem do CDR sintético dos primeiros 50 registros gerados anteriormente tem-se a listagem indicada no Quadro 39.

id usuário	id local origem	Local	Latitude	Longitude	horário inicial da chamada	duração da chamada	horário de término da chama	data da ligação
1	BTC1	Barramares Flat	-22.997283	-43.358127	0:09:55	0:03:00	0:12:55	11-06-15
1	BTC1	Barramares Flat	-22.997283	-43.358127	0:38:18	0:03:00	0:41:18	10-06-15
1	BTC1	Barramares Flat	-22.997283	-43.358127	18:10:29	0:03:00	18:13:29	08-06-15
1	BTC1	Barramares Flat	-22.997283	-43.358127	19:00:05	0:03:00	19:03:05	10-06-15
1	BTC1	Barramares Flat	-22.997283	-43.358127	19:01:15	0:03:00	19:04:15	12-06-15
1	BTC1	Barramares Flat	-22.997283	-43.358127	20:25:17	0:03:00	20:28:17	09-06-15
1	BTC1	Barramares Flat	-22.997283	-43.358127	20:54:12	0:03:00	20:57:12	10-06-15
1	BTC1	Barramares Flat	-22.997283	-43.358127	21:12:35	0:03:00	21:15:35	11-06-15
1	BTC1	Barramares Flat	-22.997283	-43.358127	22:24:25	0:03:00	22:27:25	11-06-15
1	BTC1	Barramares Flat	-22.997283	-43.358127	23:57:14	0:03:00	0:00:14	08-06-15
1	BFT1	Shoping Rio Sul	-22.95693	-43.176703	9:25:21	0:03:00	9:28:21	09-06-15
1	BFT1	Shoping Rio Sul	-22.95693	-43.176703	10:40:19	0:03:00	10:43:19	09-06-15
1	BFT1	Shoping Rio Sul	-22.95693	-43.176703	12:01:46	0:03:00	12:04:46	12-06-15
1	BFT1	Shoping Rio Sul	-22.95693	-43.176703	12:04:56	0:03:00	12:07:56	09-06-15
1	BFT1	Shoping Rio Sul	-22.95693	-43.176703	12:18:27	0:03:00	12:21:27	11-06-15
1	BFT1	Shoping Rio Sul	-22.95693	-43.176703	12:35:02	0:03:00	12:38:02	10-06-15
1	BFT1	Shoping Rio Sul	-22.95693	-43.176703	13:05:22	0:03:00	13:08:22	08-06-15
1	BFT1	Shoping Rio Sul	-22.95693	-43.176703	14:37:34	0:03:00	14:40:34	08-06-15
1	BFT1	Shoping Rio Sul	-22.95693	-43.176703	15:59:01	0:03:00	16:02:01	10-06-15
1	BFT1	Shoping Rio Sul	-22.95693	-43.176703	16:36:15	0:03:00	16:39:15	08-06-15
2	CEC2	Arcos da Lapa	-22.912919	-43.179967	6:12:00	0:03:00	6:15:00	08-06-15
2	CEC2	Arcos da Lapa	-22.912919	-43.179967	18:12:49	0:03:00	18:15:49	12-06-15
2	CEC2	Arcos da Lapa	-22.912919	-43.179967	18:37:25	0:03:00	18:40:25	10-06-15
2	CEC2	Arcos da Lapa	-22.912919	-43.179967	19:22:56	0:03:00	19:25:56	10-06-15
2	CEC2	Arcos da Lapa	-22.912919	-43.179967	20:42:30	0:03:00	20:45:30	12-06-15
2	CEC2	Arcos da Lapa	-22.912919	-43.179967	22:33:21	0:03:00	22:36:21	12-06-15
2	CEC2	Arcos da Lapa	-22.912919	-43.179967	23:09:22	0:03:00	23:12:22	10-06-15
2	BTT2	Barra Shopping	-22.997283	-43.358127	10:04:07	0:03:00	10:07:07	12-06-15
2	BTT2	Barra Shopping	-22.997283	-43.358127	10:25:54	0:03:00	10:28:54	10-06-15
2	BTT2	Barra Shopping	-22.997283	-43.358127	11:16:43	0:03:00	11:19:43	12-06-15
2	BTT2	Barra Shopping	-22.997283	-43.358127	11:36:12	0:03:00	11:39:12	11-06-15
2	BTT2	Barra Shopping	-22.997283	-43.358127	12:23:14	0:03:00	12:26:14	12-06-15
2	BTT2	Barra Shopping	-22.997283	-43.358127	14:49:34	0:03:00	14:52:34	12-06-15
2	BTT2	Barra Shopping	-22.997283	-43.358127	15:02:35	0:03:00	15:05:35	10-06-15
2	BTT2	Barra Shopping	-22.997283	-43.358127	15:04:19	0:03:00	15:07:19	10-06-15
2	BTT2	Barra Shopping	-22.997283	-43.358127	17:06:20	0:03:00	17:09:20	11-06-15
2	BTT2	Barra Shopping	-22.997283	-43.358127	17:51:24	0:03:00	17:54:24	08-06-15
3	SCT3	Centro de Tradições Nordestinas - Luiz Gonzaga	-22.89716	-43.222113	9:18:29	0:03:00	9:21:29	10-06-15
3	SCT3	Centro de Tradições Nordestinas - Luiz Gonzaga	-22.89716	-43.222113	12:51:25	0:03:00	12:54:25	12-06-15
3	SCT3	Centro de Tradições Nordestinas - Luiz Gonzaga	-22.89716	-43.222113	13:07:53	0:03:00	13:10:53	10-06-15
3	SCT3	Centro de Tradições Nordestinas - Luiz Gonzaga	-22.89716	-43.222113	13:17:32	0:03:00	13:20:32	11-06-15
3	SCT3	Centro de Tradições Nordestinas - Luiz Gonzaga	-22.89716	-43.222113	13:32:00	0:03:00	13:35:00	11-06-15
3	SCT3	Centro de Tradições Nordestinas - Luiz Gonzaga	-22.89716	-43.222113	15:16:32	0:03:00	15:19:32	10-06-15
3	SCT3	Centro de Tradições Nordestinas - Luiz Gonzaga	-22.89716	-43.222113	15:18:02	0:03:00	15:21:02	10-06-15
3	SCT3	Centro de Tradições Nordestinas - Luiz Gonzaga	-22.89716	-43.222113	15:29:21	0:03:00	15:32:21	08-06-15
3	SCT3	Centro de Tradições Nordestinas - Luiz Gonzaga	-22.89716	-43.222113	15:32:38	0:03:00	15:35:38	09-06-15
3	SCT3	Centro de Tradições Nordestinas - Luiz Gonzaga	-22.89716	-43.222113	16:06:11	0:03:00	16:09:11	09-06-15
3	SCT3	Centro de Tradições Nordestinas - Luiz Gonzaga	-22.89716	-43.222113	16:15:09	0:03:00	16:18:09	08-06-15
3	SCT3	Centro de Tradições Nordestinas - Luiz Gonzaga	-22.89716	-43.222113	16:36:27	0:03:00	16:39:27	08-06-15
3	SCC3	Condomínio Quinta do Conde	-22.906157	-43.220037	20:43:57	0:03:00	20:46:57	09-06-15

Quadro 39- *Layout* de CDR enriquecido com o posicionamento geográfico das torres de celulares

Fonte: Elaborado pelo autor

Diante desse novo *layout* de CDR sintético, foi possível disponibilizar um mapa geográfico (Figura 10) com os 50 primeiros registros de CDR sobre a cidade do Rio de Janeiro para mostrar graficamente os horários que as pessoas realizaram ligações telefônicas.

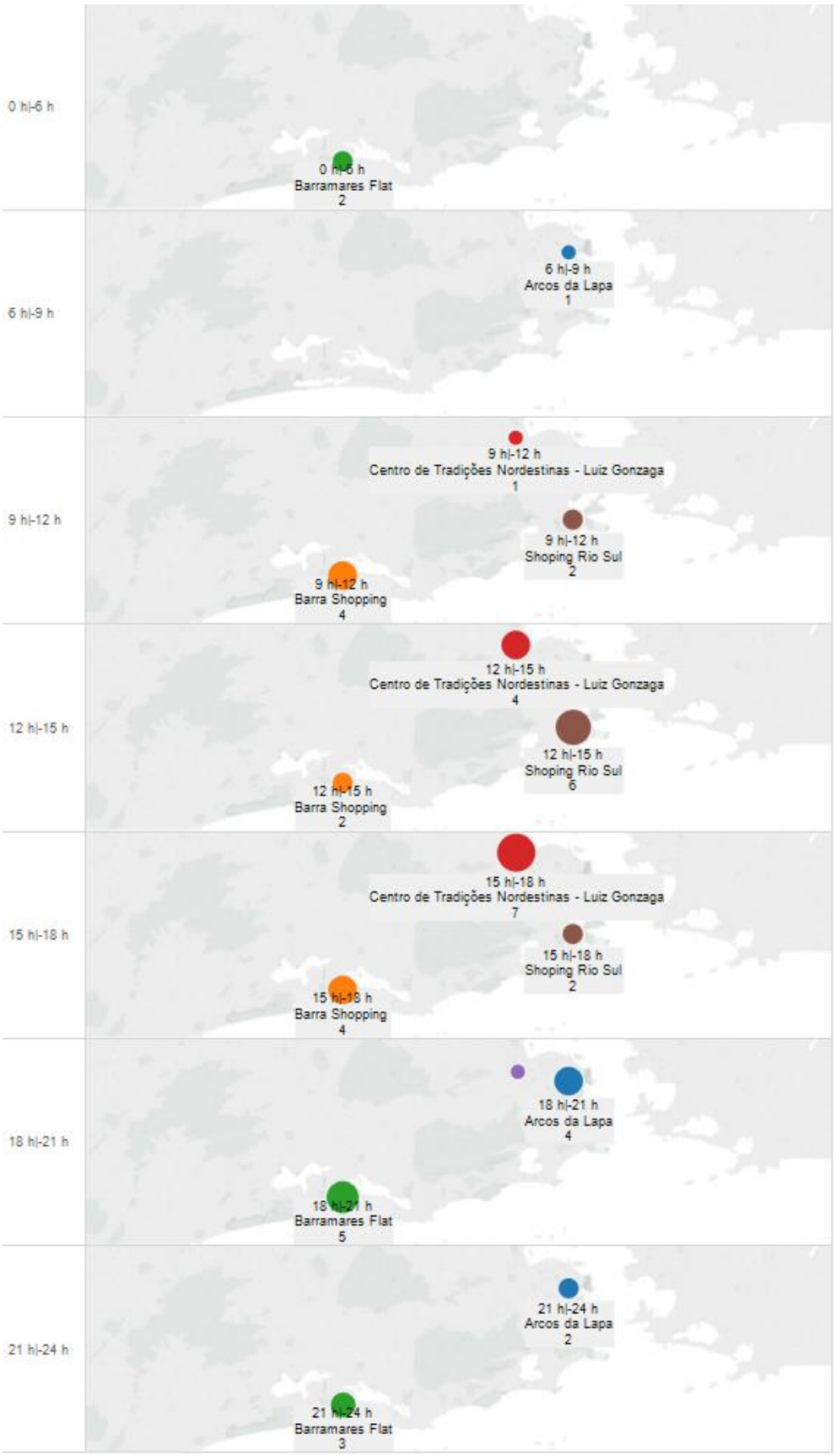


Figura 10 - Quantidade de ligações por período do dia
Fonte: Elaborado pelo autor

A Figura 10 apresenta o volume de ligações por período de horário dos primeiros três usuários da lista dos 50 primeiros registros de CDRs sintéticos gerados pelo método adaptado do professor Isaacman (2012), conforme descrito neste estudo. No período das 0h às 6h, foram realizadas duas ligações no Barramares Flat (no bairro da Barra da Tijuca). No período das 6h às 9h, foi realizada somente uma ligação nos Arcos da Lapa (no bairro do centro do Rio). No período das 9h às 12h, foram realizadas sete ligações no total, sendo: uma ligação no Centro de Tradição Nordestina – Luiz Gonzaga (no bairro de São Cristóvão), duas ligações no Shopping Rio Sul (no bairro de Botafogo) e quatro ligações no Barra Shopping (no bairro da Barra da Tijuca). No período das 12h às 15h, foram realizadas doze ligações no total, sendo elas: quatro no Centro de Tradições Nordestinas – Luiz Gonzaga (no bairro de São Cristóvão), seis ligações no Shopping Rio Sul (no bairro de Botafogo) e duas ligações no Barra Shopping (no bairro da Barra da Tijuca). No horário das 15h às 18h foram realizadas treze ligações no total, sendo elas: sete ligações no Centro de Tradição Nordestina – Luiz Gonzaga (no bairro de São Cristóvão), duas ligações no Shopping Rio Sul (no bairro de Botafogo) e quatro ligações no Barra Shopping (no bairro da Barra da Tijuca). No horário das 18h às 21h, foram realizadas nove ligações, sendo elas: quatro nos Arcos da Lapa (no bairro do Centro do Rio) e cinco no Barramares Flat (no bairro da Barra da Tijuca) e, concluindo, no horário das 21h até às 0h, foram efetuadas duas ligações nos Arcos da Lapa (no bairro do Centro do Rio).

A Figura 10 foi consolidada por número de ligações realizadas por horários. Poder-se-ia consolidar de modo diferente, como, por exemplo: por número de ligações realizadas por horários e pelos dias da semana. A importância desta informação é demonstrar a real aplicação do método descrito nesta dissertação para criar CDRs sintéticos. Identificar o fluxo de ligações telefônicas e poder segregar essas ligações por horários (faixa de horário do período do dia), por dias da semana (para observar o comportamento/hábitos dos assinantes ao longo da semana) e por localidade de origem da ligação (para identificar qual bairro requer atendimento de expansão da rede de acesso móvel mais urgentemente). Desta forma, é possível criar uma visão de demanda por volume, o que é essencial para analisar a infraestrutura de torres de acesso de telecomunicações necessárias a serem provisionadas nestes lugares (ou bairros).

Outro importante benefício é poder criar cenários (situações hipotéticas), ou melhor, cenários sintéticos, e estudar (sobre esses cenários sintéticos) o quanto o aumento de tráfego (ou mudança de perfil de tráfego) impactaria na volumetria da rede

de telecomunicações já existente. Assim, fica mais fácil planejar o crescimento da rede de acesso móvel para o crescimento da uma região qualquer, ou ainda, poder atender à demanda de megaeventos como o Rock in Rio, Olimpíadas, Copa das Confederações, Copa do Mundo, *shows* de Rock – entre outros grandes eventos de entretenimento na cidade.

Foi utilizado o *software Tableau Desktop* para gerar o mapa geoposicionado das torres de celulares móveis. Os campos utilizados foram Id usuário, horário inicial da ligação, data da ligação e a posição geográfica de onde a ligação foi completada. Somente foi possível construir a Figura 10 graças ao enriquecimento do CDR sintético com os campos de latitude, longitude e a localização de onde a ligação foi realizada.

Os campos utilizados para gerar o mapa das ligações realizadas pelos os usuários sintéticos (Figura 10) foram: Data da ligação (DD,MM,AA), duração da chamada (hh:mm.ss), horário inicial da chamada (hh:mm:ss), id do usuário (que pode ser o número de telefone sintético), local (identificação de onde a ligação foi realizada) e a latitude e longitude (local de posicionamento da antena da operadora de telecomunicações).

7 CONSIDERAÇÕES FINAIS

O *Call Detail Record* (CDR) gerado pelas redes de telecomunicações é uma rica fonte de informação para diversas aplicações em diferentes áreas da ciência, conforme foi descrito ao longo desta dissertação. Todavia, obter os CDRs para realizar pesquisas é um grande desafio para os pesquisadores de todo o mundo, pois o CDR traz consigo dados que identificam as pessoas, bem como revela os hábitos das pessoas em seu cotidiano. Por este motivo, o acesso aos CDRs é restrito por lei e, quando liberados, é por meio de liminar judicial ou por contrato de confidencialidade (NDA). Essa é uma enorme barreira para o desenvolvimento de estudos mais amplos sobre o entendimento do comportamento das pessoas, do deslocamento urbano e do planejamento de rede de acesso móvel. Simular o volume de tráfego de voz e de dados em uma rede de acesso móvel para atendimento às demandas de megaeventos, como, por exemplo, Rock in Rio, Olimpíadas, Copa das Confederações, Copa do Mundo, é atualmente um desafio. Em estudos mais recentes, como descrito ao longo desta dissertação, foi possível identificar, com o auxílio de CDR, o deslocamento urbano das pessoas portadoras do vírus do Ebola sem elas saberem. Isso foi relevante para a sociedade para entender onde as pessoas estavam originalmente e para onde elas se deslocaram posteriormente. Dessa forma, os órgãos públicos de saúde conseguiram planejar o atendimento médico para as regiões que recebiam essas pessoas que estavam contaminadas. Outros exemplos práticos de aplicabilidade do uso de CDR são a evacuação/resgate de pessoas em áreas que sofreram catástrofes naturais (identificando os últimos registros de CDRs de ligações efetuadas nessas áreas) e planejamento de cidades. Tudo isso é possível porque as pessoas realizam ligações telefônicas ao longo de seus trajetos, o que dispara, automaticamente, a geração de CDR nas centrais telefônicas e, a partir desse momento, se obtém a localização geográfica de onde as pessoas estavam.

Para os pesquisadores, não ter acesso facilitado aos CDRs tornou-se um grande obstáculo, frente ao dilema de como extrair informações úteis de CDR e ainda garantir a privacidade das pessoas, evitando o vazamento de informações sensíveis. Segundo autores vistos no decorrer desta pesquisa, a privacidade é fundamental e diz respeito à manutenção de um espaço pessoal, sem interferências de outros indivíduos ou instituições. Para minimizar riscos e aumentar a confiança da segurança do acesso aos CDRs, a alternativa é utilizar a técnica de anonimização. O termo anonimato vem do

adjetivo "anônimo" e representa o fato de o sujeito não ser unicamente caracterizado dentro de um conjunto de sujeitos. Neste caso, afirma-se que o conjunto está anonimizado. De forma análoga, o termo usado em inglês – *anonymizer* –, diz respeito à remoção das informações de identificação pessoal de cada registro. Por este motivo, os pesquisadores, atualmente, substituem os números de telefones das pessoas dos campos dos CDRs por valores numéricos sequenciais.

Zang e Bolot (2011) apresentaram à comunidade acadêmica um argumento contrário à confiabilidade do processo de anonimização e levantaram o seguinte questionamento: Deve-se ou não confiar na privacidade de publicação dos dados anonimizados de CDR? Até então, nenhum estudo havia sido realizado questionando o processo de anonimização dos campos do CDR. Mas, diante dos resultados da pesquisa que envolveu 30 bilhões de registros telefônicos de chamada de voz e contemplou 25 milhões de usuários de telefones móveis de uma operadora nacional dos EUA, abrangendo 50 estados americanos, a posição foi estarrecedora. A conclusão do estudo de Zang e Bolot (2011) foi de que a técnica de anonimização de dados de CDR não funciona. Essa posição foi ratificada pelo relatório do MIT e pela Universidade de Louvain na Bélgica, que concluíram que 95% dos usuários de telefones celulares, mesmo utilizando a técnica de anonimização, podem ser identificados com base em seus padrões de comportamento (MONTJOYE et al., 2013). Os pesquisadores concluíram, com esse estudo, que a preservação de dados anônimos não é necessariamente suficiente para garantir a privacidade real.

Para enfrentar o desafio de “como trabalhar ou manipular os dados de CDR, com toda essa riqueza de informação, sem violar a privacidade das pessoas” em pesquisas acadêmicas e na sociedade, esta dissertação apresentou o método prático da geração de CDR sintético, que permite preservar a identidade das pessoas.

Outro problema ainda apresentado nesta dissertação foram as dificuldades que as operadoras de telecomunicações enfrentam em seu dia a dia – qualidade da infraestrutura de acesso à rede de telefonia pelos usuários –, tais como: ligações telefônicas não completadas, linhas sem sinal (ou mudas), ruídos nas ligações telefônicas, interrupção inesperada das ligações e a ausência de cobertura de sinal. Esses são os principais motivos para as reclamações dos usuários das empresas de telefonia. Como foi visto na seção 2.2, estudos revelam que um dos fatores que mais influenciam os assinantes a trocarem de prestadoras de serviços de telecomunicação é a qualidade do serviço de voz e de dados. Esse fator tem importância da ordem de 21% para usuários

realizarem *churn*. Os altos índices de insatisfação registrados nos órgãos de defesa do consumidor são alarmantes. Conforme foi descrito no presente estudo, a ANATEL – órgão regulador da indústria de telecomunicações – tem realizado um bom trabalho na monitoração dos indicadores de qualidade dos serviços de telefonia no país, mas isso não tem sido suficiente para que a sociedade tenha acesso a melhores serviços de telecomunicações.

O método de geração de CDR sintético apresentado nesta dissertação é também uma ferramenta importante para o planejamento da rede de acesso das empresas de telecomunicações. A técnica utilizada permite criar cenários sintéticos (ambientes hipotéticos). Esses cenários sintéticos permitem criar volume de tráfego de dados/voz e analisar como esse volume de tráfego de carga sintética pode impactar a rede de telefonia celular real ou, ainda, como as operadoras de telecomunicações podem planejar suas expansões, identificando os melhores locais geográficos para a instalação de antenas da rede de acesso móvel. Poder utilizar o volume de tráfego de CDR sintético para gerar teste de carga e planejar melhor o crescimento das redes é atender diretamente à necessidade de qualidade de rede que tanto os consumidores desejam e, assim, reduzir o *churn*.

O CDR sintético é uma reprodução similar do CDR real e construído com base nos dados de censos geográficos públicos. O CDR sintético emprega cenários hipotéticos, mas sustentados por informações reais de pesquisas. Essa é uma alternativa para a comunidade científica realizar seus estudos, empregando CDR e preservando a identidade das pessoas. Utilizando-se o método WHERE adaptado, deixa de existir a restrição do uso de CDR, além de as operadoras de telecomunicações ganharem uma ferramenta para o planejamento, implantação e expansão de rede física. Tal método precisou ser adaptado, nesta dissertação, em relação ao método original, que foi criado pelo professor Dr. Sibren Isaacman, do Departamento de Ciência de Computação da Universidade Loyola Maryland nos EUA. O método WHERE adaptado, desenvolvido no presente estudo, traz benefícios para a academia e para a sociedade graças à ausência do uso de CDRs reais para a geração do CDR sintético (hipotético). Esta foi uma das contribuições de originalidade e de inovação introduzida nesta pesquisa, pois permitiu viabilizar a construção de CDRs sintéticos com o mesmo padrão de formatação dos CDRs originais, reproduzindo a mesma variação temporal de frequência das ligações dos CDRs originais.

Esta dissertação acrescentou ao método WHERE outros trabalhos matemáticos acadêmicos, que permitiram construir os CDRs sintéticos com o mesmo padrão de formatação dos CDRs originais. A opção de não utilizar CDRs reais como insumo para gerar os CDRs sintéticos forçou buscar uma adaptação para que os CDRs sintéticos gerados possuíssem a mesma variação temporal de frequência das ligações dos CDRs originais. O primeiro trabalho que contribuiu foi o de Queijo e Almeida (1998), sobre modelos para distribuições espaciais e temporais de tráfego em redes GSM. A parte utilizada foi a definição da densidade de tráfego – que é dada pelo modelo de variação espacial (amplitude) – e a forma envolve uma dupla gaussiana, modelando a densidade de probabilidade de se efetuar uma ligação ao longo de um dia. Adicionalmente, o trabalho de Olver e Towsend (2006) foi essencial para viabilizar o sorteio dos horários das ligações de cada usuário de acordo com curva de probabilidade que define o padrão de ligação de uma determinada população. Os sorteios foram realizados sobre um espaço não equiprovável, modelado por uma função matemática qualquer com variáveis aleatórias comuns, através de inversão de função de probabilidade acumulada. Esses dois trabalhos acadêmicos, inseridos nesta dissertação, viabilizaram a geração de CDR sintético, o que tornou o resultado deste trabalho original e inovador.

O método de geração de CDRs sintéticos, desenvolvido pelo professor Dr. Sibren Isaacman, é inovador (por ser o único trabalho acadêmico que descreve como criar um CDR), além de ser totalmente flexível e passível de adaptações e enriquecimentos, aproximando-se ainda mais dos CDRs reais. Ao longo desta dissertação, além do método original (citado anteriormente) ter sido adaptado, o conteúdo dos campos de CDR foi também enriquecido em relação ao projeto inicial do método WHERE. Foram inseridos os campos latitude e longitude (localização geográfica das torres de celulares da operadora) identificando o local, dentro do bairro, no qual as ligações telefônicas foram realizadas. Devido a esse enriquecimento no CDR sintético, foi possível utilizar os mesmos *softwares* (programas de computadores) que são utilizados pelos CDRs reais para apresentar um mapa com o volume de ligações por período de horário e a posição geográfica de onde as ligações foram realizadas.

A importância e as diversas aplicabilidades para o uso de CDR tanto para as operadoras de telecomunicações quanto para a comunidade acadêmica e para a sociedade foi amplamente debatida neste trabalho. A revisão da literatura discutiu os diferentes exemplos de uso do CDR em diferentes campos da ciência. Não há dúvida sobre o quanto é importante investigar o movimento das “coisas”, através do CDR. O

CDR é nativo do ambiente das telecomunicações. Segundo o relatório *Global System Mobile Association* (GSMA, 2015), o mercado de telefonia móvel continua crescendo rapidamente. No final de 2014, o total de assinantes móveis atingiu 3,6 bilhões de usuários, metade da população mundial. A previsão para 2020 é atingir mais um bilhão de assinantes, chegando a 60% de penetração mundial. Outro importante segmento do mercado de telecomunicações, e que possui enorme potencial de crescimento, é o mercado M2M (*machine to machine*), também chamado de internet das coisas (IOT). Segundo o mesmo relatório, no final de 2014, 243 milhões de celulares estavam conectados ao M2M, com previsão de crescimento anual de 26% de 2014 até 2020. Todo esse potencial mercadológico amplia a importância desta dissertação no aspecto da abordagem do uso do método WHERE, desenvolvido pelo professor Dr. Sibren Issacman.

Neste trabalho, foi realizada a prática da produção de um conjunto de CDR sintético e apresentadas as evidências de que o CDR sintético pode ser muito útil para os estudos de planejamento, de implantação e da expansão da rede de telefonia móvel, pela sua capacidade de simular e prever os impactos de mudanças nas redes de telecomunicações. Os quatro grandes principais diferenciais do método WHERE são: (i) produzir CDR sintético com características bem similares aos CDRs reais; (ii) a relação da privacidade, conforme exposto ao longo desta dissertação; (iii) a técnica do CDR sintético pode se tornar uma ferramenta valiosa para as operadoras; (iv) utilização do método WHERE para estudar e inferir na mecânica da mobilidade urbana em grandes centros urbanos.

Ao analisar cada um desses quatro itens, é possível depreender que, quanto ao item (i), a produção de CDR sintético com características bem similares aos CDRs reais, torna, praticamente, desnecessário o uso de CDRs reais, produzidos pelos elementos de redes das operadoras de telecomunicações. No que se refere ao item (ii), a privacidade, conforme exposto ao longo desta dissertação, fica garantida em estudos que utilizem os CDRs sintéticos. Quanto ao item (iii), a técnica do CDR sintético pode se tornar uma ferramenta para as operadoras planejarem a expansão das suas redes de acessos móveis, simulando, em laboratório, cenários “sintéticos”, como cidades “sintéticas”, população “sintética” e o volume de carga de tráfego também “sintético”, introduzindo essas perturbações (situações de cargas específicas) nesses cenários para observar novos comportamentos da rede de telefonia. O item (iv) diz respeito à utilização do método WHERE para estudar e inferir na mecânica da mobilidade urbana em grandes centros

urbanos, o que pode contribuir para a melhoria de políticas públicas de forma integrada com o desenvolvimento regional e o avanço tecnológico.

Esses quatro fatores permitem que a comunidade acadêmica, outras empresas de pesquisa e entidades governamentais possam fazer estudos de impactos de crescimento das cidades, deslocamento urbano das pessoas, sem preocupação em relação à possível revelação da identidade das pessoas. Por este motivo, esta dissertação concentrou-se na produção de CDR sintético para a realização de simulação de tráfego de rede de telefonia e estudo de mobilidade urbana, mostrando o quanto o método WHERE atende e resolve a preocupação e a preservação da identidade das pessoas.

O objetivo desta dissertação foi o de apresentar uma técnica que permita realizar pesquisas com CDR, preservando a privacidade das pessoas – o que foi plenamente atendido, utilizando-se fontes oficiais, como dados atuários divulgados por instituto de pesquisa brasileiro. Em função das diferentes aplicabilidades apresentadas nesta dissertação sobre o uso de CDR, mostra-se essencial a intensificação de estudos de simulação de tráfego nas redes de telefonia móvel, bem como aprofundamento de pesquisas em mobilidade urbana, fazendo uso de CDR sintético.

Recomendação para trabalhos futuros

O desdobramento desta pesquisa estará associado na aplicabilidade de CDRs sintéticos de soluções práticas no deslocamento urbano de pessoas e planejamento de expansão de rede de telecomunicações móveis, por exemplo:

- Integração com as torres de celulares disponíveis no site Opencellid => http://opencellid.org
- Inclusão de outros lugares, além de casa/trabalho
- Inclusão do tempo de transporte por diferentes meios
- Incluir informações públicas como dados SMTU (Secretaria Municipal de Transporte Urbano).
- Adaptação do método para contemplar mais a mobilidade urbana ou mais planejamento de crescimento de rede de telefonia.

Estar associado a um programa de pós-graduação é mister[82] para criar um grupo de pesquisa sobre o uso de CDR sintético nos estudos acadêmicos e corporativos.

[82] Condição necessária ou de exigência.

8 REFERÊNCIAS BIBLIOGRÁFICAS

AGÊNCIA NACIONAL DE TELECOMUNICAÇÕES. ANATEL. Relatório de Indicadores de Desempenho Operacional. 2014. Disponível em: <http://www.anatel.gov.br/Portal/verificaDocumentos/documento.asp?numeroPublicacao=331461&filtro=1&documentoPath=331461.pdf>. Acesso em: 13 set.2015.

AGINSKY, A. **Is big data the next big thing for telecom?** EMEA, n. 12, 2012. Disponível em: <http://www.connect-world.com/index.php/magazines/emea/item/18331-is-big-data-the-next-big-thing-for-telecom>. Acesso em: 21 nov. 2014.

BARABÁSI, A. L.; GONZÁLEZ, M. C.; HIDALGO, C. A. Understanding individual human mobility patterns. **Nature Publishing Group**, v. 453, June, 2008.

BECKER, R. C. R. et al. Route classification using cellular handoff patterns. In: 13th International Conference on Ubiquitous Computing (Ubicomp), Sept. 2011.

BOOTH, W. et al. The craft of research. Chicago: University of Chicago, 3rd edition, 2008.

BRANCO JUNIOR, E. C.; MACHADO, J. C.; MONTEIRO, J. M. Estratégias para Proteção da privacidade de dados armazenados na nuvem. SBC 1ª Ed. ISBN 978-85-7669-290-4, 2014. Disponível em: <http://www.inf.ufpr.br/sbbd-sbsc2014/sbbd/proceedings/artigos/pdfs/14.pdf>. Acesso em: 23 jan. 2016.

BRASIL. **Lei nº 12.965**, de 23 de abril de 2014. Estabelece princípios, garantias, direitos e deveres para o uso da internet no Brasil. Diário Oficial da União, ano CLI nº 77, Brasília, 9 de abril de 2014.

CANDIA, M. C. et al. Uncovering individual and collective human dynamics from mobile phone records. MATH.THEOR., 41:224015, 2008.

CLARKE, R. Introduction to dataveillance and information privacy, and definition

COSTA, S. R. **Análise estatística na conciliação de receita de público e despesa de uso de redes em operadora de Telecom**. 2010. Dissertação (Mestrado em Engenharia Elétrica)— Departamento de Engenharia Elétrica, Universidade Nacional de Brasília, Distrito Federal, Brasil, 2010.

CUKIERMAN, H. L. **Yes, nós temos Pasteur**: Manguinhos, Oswaldo Cruz e a história da ciência no Brasil. Rio de Janeiro: Ed. Relume Dumará, FAPERJ, 2007.

DALFOVO, M. S.; LANA, R. A.; SILVEIRA, A. Métodos quantitativos e qualitativos: um resgate teórico. **Revista Interdisciplinar Científica Aplicada**. Blumenau, v. 2, n. 4, p.01-13, Sem II. 2008.

FARIA, G. F. **Análise de padrões em chamadas telefônicas**. 2010. Relatório final da disciplina Princípios a Aplicações de Mineração de Dados (CAP-359) do Programa de Pós-Graduação em Computação Aplicada do Instituto Nacional de Pesquisas Espaciais (INPE). Disponível em: <http://urlib.net/8JMKD3MGP7W/38FUMP2>. Acesso em: 11 mar. 2015.

FONSECA, M.V.A. **Gestão da inovação**: elos de valor do ambiente 21. Apostila da disciplina Inovação nas Organizações do Programa de Engenharia de Produção da COPPE. Rio de Janeiro: Universidade Federal Federal do Rio de Janeiro, 2013.

GONZÁLEZ, M. C.; HIDALGO, C. A.; BARABÁSI, A.-L.. Understanding individual human mobility patterns. **Nature**, v. 453, p. 779-782, 2008.

GSMA. **The mobile economy**. Global System Mobile Association, 2015. Disponível em: <http://www.gsmamobileeconomy.com/GSMA_Global_Mobile_Economy_Report_2015.pdf>. Acesso em: 21 jun. 2015.

HANSON, K. et al. A tale of one city: using cellular network data for urban planning. IEEE Pervasive Computing, v. 10, n. 4, Oct./Dec. 2011.

HU, H. et al. Toward scalable systems for big data analytics: a technology tutorial, IEEE Access, v. 2, p 652-687, July 2014. Disponível em <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6842585>. Acesso em: 02 out. 2015.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. IBGE. **Informações sobre bairros segundo os municípios**: Censo 2010. Disponível em: <www.ibge.gov.br/home/presidencia/noticias/imprensa/ppts/0000000486.xls&sa=U&ei=7mghVfyDEYGjsQXUm4HQCw&ved=0CAQQFjAB&client=internal-uds-cse&usg=AFQjCNFaZXyT7tsv3OBo2PAL8_UksluF2Q>. Acesso em: 05 abr. 2015.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. IBGE. **Informações sobre bairros segundo os municípios**: reflexões sobre os deslocamentos populacionais no Brasil 2011. Disponível em: <http://biblioteca.ibge.gov.br/visualizacao/livros/liv49781.pdf>. Acesso em: 15 set. 2015.

INSTITUTO DE PESQUISA ECONÔMICA APLICADA**.** IPEA. **Mobilidade urbana 2011**. Disponível em: <http://www.ipea.gov.br/portal/images/stories/PDFs/SIPS/110504_sips_mobilidadeurbana.pdf>. Acesso em: 15 set. 2015.

INSTITUTO MUNICIPAL DE URBANISMO PEREIRA PASSOS. Secretaria Extraordinária de Desenvolvimento da Prefeitura do Rio de Janeiro, 2008. Disponível em: <http://portalgeo.rio.rj.gov.br/estudoscariocas/download%5C2938_Distribui%C3%A7%C3%A3o%20dos%20empregos%20na%20cidade%20do%20Rio%20de%20Janeiro%20em%202008.pdf>. Acesso em: 11 abr. 2015.

ISAACMAN, S. et al. A tale of two cities. In: Workshop on Mobile Computing Systems and Applications (HotMobile), Feb. 2010.

ISAACMAN, S. et al. Human mobility modeling at metropolitan scales. In: International Conference on Mobile Systems, X., 2012, Low Wood Bay. **Anais…** Low Wood Bay, Lake District, UK: MobiSys'12, 2012.

ISAACMAN, S. **Modeling the impact of human mobility**: mobile devices as sensors and content vectors, Doctoral Thesis, Princeton University, NJ, USA. June, 2012.

KOEN, B. V. **Discussion of the method**: conducting the engineer's approach to problem solving. New York: Oxford University Press, 2003.

LAKATOS, I. **Philosophical papers**: the methodology of scientific research programmes. Volume I. Cambridge: Cambridge University Press, 1999.

LENHARO, M. **Dados de celulares podem ajudar no combate à epidemia de ebola**. G1, 30 out. 2014. Disponível em: <http://g1.globo.com/bemestar/ebola/noticia/2014/10/dados-de-celulares-podem-ajudar-no-combate-epidemia-de-ebola.html>. Acesso em: 01 maio 2015.

MARQUES NETO, H. T. et al. Análise da mobilidade humana em eventos de larga escala baseada em chamadas de telefones celulares. In: Congresso da Sociedade Brasileira de Computação, XXXIII., 2013, Maceió. **Anais…** Maceió: CSBC. 2013.

MEDEIROS, F.[experiência com CDR realizada no Réveillon de 2013/2014 em Copacabana]. Rio de Janeiro, 2014. Entrevista de Pablo Cerqueira, Cientista de Dados no Centro de Operações Rio. 27 de outubro de 2014.

MICHAEL, C. M. et al. Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. In: IEEE Transactions on Neural Networks. Department of Computer Science University of Colorado Boulder, 2000.

MIHESSEN, V.; MACHADO, D. C.; PERO, V. Mobilidade urbana e mercado de trabalho na Região Metropolitana do Rio de Janeiro. In: Encontro Nacional de Economia, XLII., 2014, Natal. **Anais...** Natal: ANPEC. Disponível em: <http://www.anpec.org.br/encontro/2014/submissao/files_I/i10-1dc14346dd67760748fefecaac00a05a.pdf>. Acesso em: 07 abr. 2015.

MONTJOYE, Y-A. et al. Unique in the crowd: the privacy bounds of human mobility, **Nature Scientific Reports**. v. 3, March, p. 1-5, 2013.

of terms, 1999. Disponível em: <http://www.qatar.cmu.edu/iliano/courses/10F-CMU-CS349/slides/privacy.pdf >. Acesso em: 23 jan. 2016.

OLVER, S.; TOWNSEND, A. **Fast inverse transform sampling in one and two dimensions**. School of Mathematics and Statistics, The University of Sydney, Australia, 2006. Disponível em: <http://www.maths.usyd.edu.au/u/olver/papers/InverseTransformSampling.pdf>. Acesso em: 13 jun. 2015.

PLATÃO, **A República**, trans. Carlos Alberto Nunes. Belém: Universidade Federal do Pará, 2000.

QUEIJO, J.; ALMEIDA, S. **Modelos para distribuições espaciais e temporais de tráfego em GSM**. Secção de Propagação e Radiação. Departamento de Engenharia Electrotécnica e Computadores, Lisboa: Instituto Superior Técnico. Setembro, 1998.

SÁNCHEZ, D. F. M. Mobilidade e transporte público na cidade contemporânea, o caso da cidade metropolitana de Quito - Equador. Rio de Janeiro: UFRJ/FAU, 2013.

SILVA, E. L. **Metodologia da pesquisa e elaboração de dissertação**. SILVA, E. L.;, MENEZES, E. M. (Org.) 4. ed. rev. atual. Florianópolis: UFSC, 2005.

SILVA, J. C. Caracterização e análise do deslocamento "casa-trabalho-casa" em empresas localizadas na Barra da Tijuca – RJ. 2014. Dissertação (Mestrado em Engenharia de Transportes) — Programa de Pós-Graduação em Engenharia de Transportes, COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2014.

SONG, C. et al. Limits of predictability in human mobility. **Science**, v. 327, p. 1018-1021, 2010.

STEWART, T. A. Capital Intelectual: a nova vantagem competitiva das empresas. Rio de Janeiro: Editora Campus, 1998.

TADEU, E. [**TI INSIDE** - Converge Comunicações] Washington, 2012. Entrevista de Raphael Stein no TI Inside Online em 25 de outubro de 2012. Disponível em: <http://convergecom.com.br/tiinside/25/10/2012/oi-cria-modelo-de-metrica-para-reduzir-churn-com-base-em-redes-sociais/#.VXd9YM9Vikq>. Acesso em: 09 jun. 2015.

TEERAYUT, H. A study on urban mobility and dynamic population estimation by using aggregate mobile phone sources. In: Center for Spatial Information Science. Tokyo: Master dissertation of the University of Tokyo. 2012. Disponível em: <http://www.csis.u-tokyo.ac.jp/english/dp/dp.html>. Acesso em: 16 maio 2015.

TELECO. **Portal de informações do setor de telecomunicações do Brasil.** Sistema de bilhetagem. Disponível em: http://www.teleco.com.br/glossario.asp?termo=bilhetagem&Submit=OK. Acesso em: 08 mar. 2015.

TELECO. **Portal de informações do setor de telecomunicações do Brasil.** Estatística Brasil. Disponível em: <http://www.teleco.com.br/estatis.asp>. Acesso em: 01 ago. 2015.

TELEMANAGEMENT FORUM (TMForum). **Business process management of telecommunication companies**. 2011. Disponível em: <https://www.tmforum.org/>. Acesso em: 13 set. 2015.

TUBE, E. Conceitos básicos sobre Erlang e tráfego telefônico. In: **Teleco – Inteligência em telecomunicações.** 2003. Disponível em: <http://www.teleco.com.br/tutoriais/tutorialerlang/>. Acesso em: 19 abr. 2015.

UNIÃO EUROPEIA. Comissão emprego, crescimento e investimento: Business opportunities. 2013. Disponível em: <https://ec.europa.eu/growth/tools-databases/dem/sites/default/files/page-files/big_data_v1.1.pdf>. Acesso em: 02. Out. 2015.

WEINBERGER, D. The machine that would predict the future. **Scientific American**, v. 305, n. 6, December.p. 52-57, 2011.

WESOLOWSKI, A. et al. **Containing the ebola outbreak**: the potential and challenge of mobile network data. LOS Currents Outbreaks. 2014 Sep 29 . Edition 1. DOI: 10.1371/currents.outbreaks.0177e7fcf52217b8b634376e2f3efc5e. Disponível em: <http://currents.plos.org/outbreaks/article/containing-the-ebola-outbreak-the-potential-and-challenge-of-mobile-network-data/>. Acesso em: 02 nov. 2015.

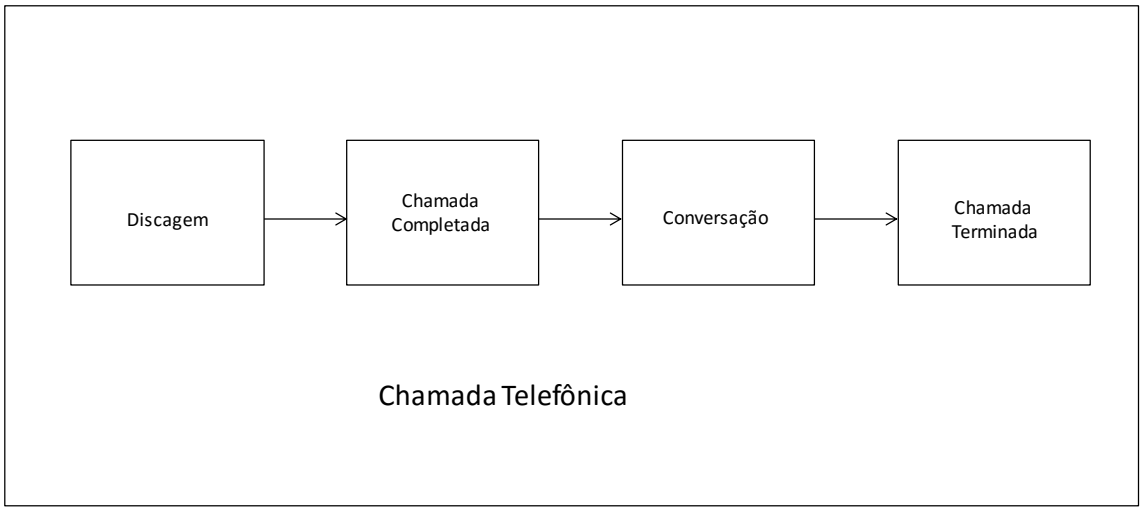
YAN, L.; FASSINO, M.; BALDASARE, P. Predicting customer behavior via calling links. In: The International Joint Conference on Neural Networks. Montreal, Canada, July, 2005.

ZANG, H.; BOLOT, J. Anonymization of location data does not work: a large-scale measurement study. In: International Conference on Mobile Computing and Networking (MobiCom), 17., 2011, New York. **Anais…** New York: ACM, 2011, p. 145-156. Disponível em: <https://mail.google.com/mail/u/0/?tab=wm#inbox/152daff68998cb9b?projector=1>. Acesso em: 02 nov. 2015.

ANEXO I

Conceitos Erlang:

O Erlang é utilizado para dimensionamento de centrais telefônicas. Segundo Tube (2003), a chamada telefônica é o processo que visa estabelecer a comunicação entre usuários utilizando dois terminais do sistema telefônico, como representado no Quadro 40.



Quadro 40 - Etapas simplificadas da chamada telefônica
Fonte: Tube, 2003

O processo inicia-se com a discagem do número telefônico com quem se deseja falar. Quando a chamada resulta em comunicação com o destino desejado, a chamada é dita completada.

As principais razões para o congestionamento na rede são:

- Congestionamento em uma das Centrais. As Centrais são dimensionadas para suportar um número máximo de tentativas de chamadas em um determinado período de tempo. O parâmetro normalmente utilizado é o *Business Hour Call Atempt* (BHCA), que equivale ao número de tentativas de chamadas na Hora de Maior Movimento (HMM).
- Congestionamento nos troncos que ligam uma central a outra. O tronco-padrão no Brasil é um circuito de 2Mbit/s (E1), com capacidade de 30 canais telefônicos (conversações).

Diante das razões apresentadas para o congestionamento telefônico, a pergunta a ser feita é: Qual o número de troncos necessários para garantir o número de troncos em um período de maior movimento? Para responder a esta questão, a fórmula desenvolvida por Agner Krarup Erlang[83] foi, e continua sendo, fundamental. Por isso, para dimensionar um sistema, é preciso estabelecer o número médio de chamadas e a duração média de cada chamada na Hora de Maior Movimento (HMM).

[83] Erlang é uma unidade de medida de intensidade de tráfego telefônico para um intervalo de uma hora.

APÊNDICE I

O Quadro 41 apresenta o sorteio dos 50 lugares de residência.

Nº do sorteio	Bairro de residência sorteado	Latitude	Longitude
1	Barra da Tijuca	Não sorteado	Não sorteado
2	Centro	Idem	Idem
3	São Cristóvão	Idem	Idem
4	Tijuca	Idem	Idem
5	Madureira	Idem	Idem
6	Copacabana	Idem	Idem
7	Copacabana	Idem	Idem
8	Copacabana	Idem	Idem
9	Madureira	Idem	Idem
10	Tijuca	Idem	Idem
11	São Cristóvão	Idem	Idem
12	Botafogo	Idem	Idem
13	Tijuca	Idem	Idem
14	Botafogo	Idem	Idem
15	Copacabana	Idem	Idem
16	Barra da Tijuca	Idem	Idem
17	Meier	Idem	Idem
18	Tijuca	Idem	Idem
19	Tijuca	Idem	Idem
20	Tijuca	Idem	Idem
21	Madureira	Idem	Idem
22	Barra da Tijuca	Idem	Idem
23	Tijuca	Idem	Idem
24	Centro	Idem	Idem
25	Barra da Tijuca	Idem	Idem
26	Botafogo	Idem	Idem
27	Botafogo	Idem	Idem
28	Madureira	Idem	Idem
29	Copacabana	Idem	Idem
30	Barra da Tijuca	Idem	Idem
31	São Cristóvão	Idem	Idem
32	Madureira	Idem	Idem
33	Centro	Idem	Idem
34	Copacabana	Idem	Idem
35	São Cristóvão	Idem	Idem
36	Copacabana	Idem	Idem
37	Barra da Tijuca	Idem	Idem
38	Copacabana	Idem	Idem
39	Copacabana	Idem	Idem
40	Tijuca	Idem	Idem

(Quadro 41 – continuação)

Nº do sorteio	Bairro de residência sorteado	Latitude	Longitude
41	Copacabana	Idem	Idem
42	Copacabana	Idem	Idem
43	Copacabana	Idem	Idem
44	São Cristóvão	Idem	Idem
45	Madureira	Idem	Idem
46	Tijuca	Idem	Idem
47	Tijuca	Idem	Idem
48	Barra da Tijuca	Idem	Idem
49	Tijuca	Idem	Idem
50	Madureira	Idem	Idem

Quadro 41 - Lista do sorteio dos 50 lugares de residência sorteados

APÊNDICE II

O Quadro 42 apresenta a lista do sorteio dos 50 lugares de trabalho.

Nº do sorteio	Bairro de trabalho sorteado	Latitude	Longitude
1	Botafogo	Não sorteado	Não sorteado
2	Tijuca	Idem	Idem
3	São Cristóvão	Idem	Idem
4	Meier	Idem	Idem
5	Tijuca	Idem	Idem
6	Centro	Idem	Idem
7	Barra da Tijuca	Idem	Idem
8	Barra da Tijuca	Idem	Idem
9	São Cristóvão	Idem	Idem
10	Centro	Idem	Idem
11	São Cristóvão	Idem	Idem
12	Botafogo	Idem	Idem
13	Barra da Tijuca	Idem	Idem
14	Botafogo	Idem	Idem
15	Centro	Idem	Idem
16	Barra da Tijuca	Idem	Idem
17	Centro	Idem	Idem
18	Copacabana	Idem	Idem
19	Botafogo	Idem	Idem
20	Centro	Idem	Idem
21	Centro	Idem	Idem
22	Centro	Idem	Idem
23	Centro	Idem	Idem
24	Barra da Tijuca	Idem	Idem
25	Centro	Idem	Idem
26	Madureira	Idem	Idem
27	Centro	Idem	Idem
28	Centro	Idem	Idem
29	Barra da Tijuca	Idem	Idem
30	Botafogo	Idem	Idem
31	Centro	Idem	Idem
32	Centro	Idem	Idem
33	Centro	Idem	Idem
34	Centro	Idem	Idem
35	Meier	Idem	Idem
36	Centro	Idem	Idem
37	Botafogo	Idem	Idem
38	Centro	Idem	Idem
39	Centro	Idem	Idem
40	Barra da Tijuca	Idem	Idem

(Quadro 42 – continuação)

Nº do sorteio	Bairro de trabalho sorteado	Latitude	Longitude
41	São Cristóvão	Idem	Idem
42	Copacabana	Idem	Idem
43	São Cristóvão	Idem	Idem
44	Centro	Idem	Idem
45	São Cristóvão	Idem	Idem
46	Madureira	Idem	Idem
47	Centro	Idem	Idem
48	Centro	Idem	Idem
49	Meier	Idem	Idem
50	Centro	Idem	Idem

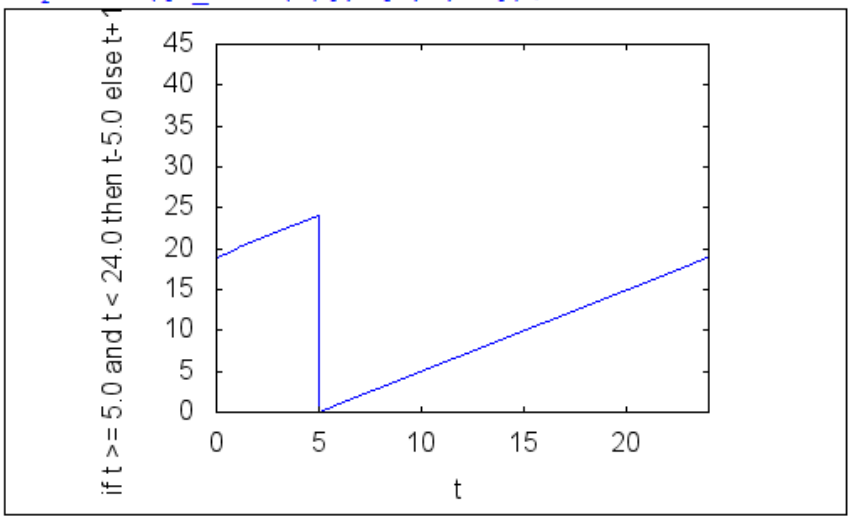
Quadro 42 - Lista do sorteio dos 50 lugares de trabalho sorteados

APÊNDICE III

Folha de trabalho gerada pelo Maxima:

```
(%i12) t_desv(2);
(%o12) 21.0
(%i13) wxplot2d([t_desv(t)], [t,0,24])$
```

(%t13)



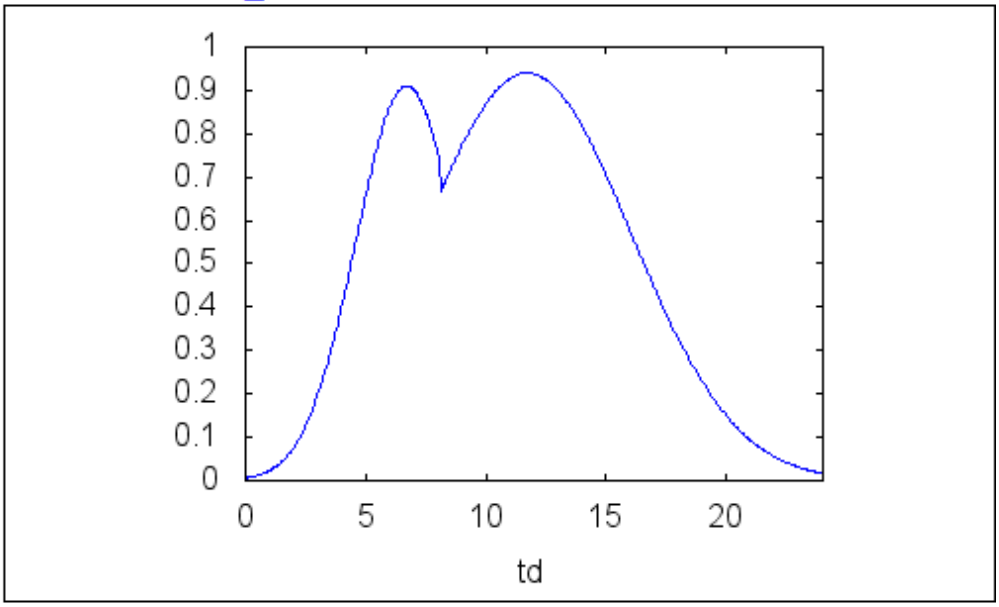

```
(%i14) %/*p1 é a ampplitude da primeira gaussiana*/;
(%o14)

p1 é a ampplitude da primeira gaussiana
h1d é a hora de pico da manhã desviada, ou seja, t_desv(h1);
d1 é o desvio da primeira gaussiana;
h_alm_d é a hora de almoço desviada, ou seja, t_desv(h_alm);
p2 é a amplitude da segunda gaussiana;
h2d é a hora de pico da tarde desviada, ou seja, t_desv(h2);
d2 é o desvio da segunda gaussiana.

(%i15) p1:0.91;
       h1:11.7;
       h1d:t_desv(h1);
       d1:2.11;
       h_alm:13.1;
       h_alm_d:t_desv(h_alm);
       p2:0.94;
       h2:16.7;
       h2d:t_desv(h2);
       d2:4.32;
(%o15) 0.91
 (%o16) 11.7
 (%o17) 6.699999999999999
 (%o18) 2.11
 (%o19) 13.1
 (%o20) 8.1
 (%o21) 0.94
 (%o22) 16.7
 (%o23) 11.7
 (%o24) 4.32
```

```
(%i39) ap_gauss(td):=if td < h_alm_d then p1*exp(-((td-h1d)^2)/(2*(d1^2)))
                     else p2*exp(-((td-h2d)^2)/(2*(d2^2)));
```

(%o39) $$\text{ap\_gauss}(td) := \text{if } td < h\_alm\_d \text{ then } p1 \exp\left(\frac{-(td-h1d)^2}{2\,d1^2}\right) \text{ else } p2 \exp\left(\frac{-(td-h2d)^2}{2\,d2^2}\right)$$

```
(%i24) wxplot2d([ap_gauss(td)], [td,0,24],[grid,10,10]);
```

(%t24)



(%o24)

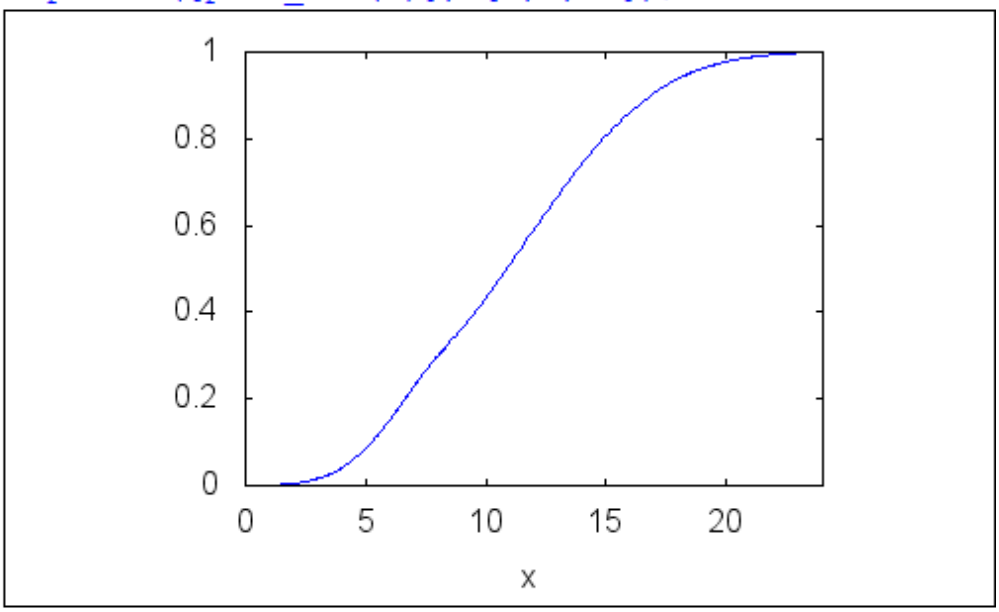
```
(%i25) quad_qags(ap_gauss(td), td, 0, 24);
(%o25) [11.68625855308805,2.953040834086096 10^-8,651,0]
(%i26) area:quad_qags(ap_gauss(td), td, 0, 24)[1];
(%o26) 11.68625855308805
(%i27) w(x):=quad_qags(ap_gauss(td), td, 0, x);
(%o27) w(x) :=quad_qags(ap_gauss(td),td,0,x)
(%i28) w(1);
(%o28) [0.01301390660498729,1.444833875317829 10^-16,21,0]
(%i29) prob_ac2(x):=w(x)[1]/area;
```

(%o29) $$\text{prob\_ac2}(x) := \frac{(w(x))_1}{area}$$

```
(%i30) prob_ac2(4)-prob_ac2(3);
(%o30) 0.02495222509163323
```

```
(%i31) wxplot2d([prob_ac2(x)], [x,0,24]);
```

(%t31)



(%o31)

```
(%i32) find_root (1/2=prob_ac2(x),x,0,24);
(%o32) 10.86925161532086

(%i33) s1: make_random_state (654321)$

(%i34) set_random_state (s1);
(%o34) done

(%i35)  random (1.0);
(%o35) 0.2260744399967471
```

APÊNDICE IV

O Quadro 43 exibe os registros completos de CDR registrados cronologicamente.

Nº do sorteio	Horário sorteado	+ 5 horas de ajuste	Horário inicial sorteado ajustado	Duração da chamada	Horário de término da chamada
1	13:10:29	5:00:00	18:10:29	00:03:00	18:13:29
2	8:05:22	5:00:00	13:05:22	00:03:00	13:08:22
3	18:57:14	5:00:00	23:57:14	00:03:00	00:00:14
4	9:37:34	5:00:00	14:37:34	00:03:00	14:40:34
5	11:36:15	5:00:00	16:36:15	00:03:00	16:39:15
6	5:40:19	5:00:00	10:40:19	00:03:00	10:43:19
7	7:04:56	5:00:00	12:04:56	00:03:00	12:07:56
8	4:25:21	5:00:00	9:25:21	00:03:00	9:28:21
9	15:25:17	5:00:00	20:25:17	00:03:00	20:28:17
10	10:59:01	5:00:00	15:59:01	00:03:00	16:02:01
11	14:00:05	5:00:00	19:00:05	00:03:00	19:03:05
12	7:35:02	5:00:00	12:35:02	00:03:00	12:38:02
13	19:38:18	5:00:00	00:38:18	00:03:00	00:41:18
14	15:54:12	5:00:00	20:54:12	00:03:00	20:57:12
15	16:12:35	5:00:00	21:12:35	00:03:00	21:15:35
16	7:18:27	5:00:00	12:18:27	00:03:00	12:21:27
17	17:24:25	5:00:00	22:24:25	00:03:00	22:27:25
18	19:09:55	5:00:00	00:09:55	00:03:00	00:12:55
19	7:01:46	5:00:00	12:01:46	00:03:00	12:04:46
20	14:01:15	5:00:00	19:01:15	00:03:00	19:04:15
21	1:12:00	5:00:00	6:12:00	00:03:00	6:15:00
22	12:51:24	5:00:00	17:51:24	00:03:00	17:54:24
23	5:25:54	5:00:00	10:25:54	00:03:00	10:28:54
24	10:04:19	5:00:00	15:04:19	00:03:00	15:07:19
25	10:02:35	5:00:00	15:02:35	00:03:00	15:05:35
26	14:22:56	5:00:00	19:22:56	00:03:00	19:25:56
27	13:37:25	5:00:00	18:37:25	00:03:00	18:40:25
28	18:09:22	5:00:00	23:09:22	00:03:00	23:12:22
29	12:06:20	5:00:00	17:06:20	00:03:00	17:09:20
30	6:36:12	5:00:00	11:36:12	00:03:00	11:39:12
31	13:12:49	5:00:00	18:12:49	00:03:00	18:15:49
32	6:16:43	5:00:00	11:16:43	00:03:00	11:19:43
33	7:23:14	5:00:00	12:23:14	00:03:00	12:26:14
34	9:49:34	5:00:00	14:49:34	00:03:00	14:52:34
35	17:33:21	5:00:00	22:33:21	00:03:00	22:36:21
36	15:42:30	5:00:00	20:42:30	00:03:00	20:45:30
37	5:04:07	5:00:00	10:04:07	00:03:00	10:07:07
38	11:36:27	5:00:00	16:36:27	00:03:00	16:39:27

(Quadro 43 – continuação)

Nº do sorteio	Horário sorteado	+ 5 horas de ajuste	Horário inicial sorteado ajustado	Duração da chamada	Horário de término da chamada
39	11:15:09	5:00:00	16:15:09	00:03:00	16:18:09
40	10:29:21	5:00:00	15:29:21	00:03:00	15:32:21
41	10:32:38	5:00:00	15:32:38	00:03:00	15:35:38
42	11:06:11	5:00:00	16:06:11	00:03:00	16:09:11
43	15:43:57	5:00:00	20:43:57	00:03:00	20:46:57
44	10:18:02	5:00:00	15:18:02	00:03:00	15:21:02
45	8:07:53	5:00:00	13:07:53	00:03:00	13:10:53
46	4:18:29	5:00:00	9:18:29	00:03:00	9:21:29
47	10:16:32	5:00:00	15:16:32	00:03:00	15:19:32
48	8:32:00	5:00:00	13:32:00	00:03:00	13:35:00
49	8:17:32	5:00:00	13:17:32	00:03:00	13:20:32
50	7:51:25	5:00:00	12:51:25	00:03:00	12:54:25

Quadro 43 - Lista de horários sorteados

APÊNDICE V

O Quadro 44 exibe as ligações sorteadas segundo função degrau.

Quantidade de ligações realizadas (formato decimal)	Quantidade de ligações realizadas (formato horário h/mm/ss)
8.33	8:19:36
15.36	15:21:40
9.54	9:32:26
22.18	22:10:40
10.43	10:25:44
13.21	13:12:24
6.52	6:31:15
7.94	7:56:30
2.58	2:34:30
17.64	17:38:16
12.26	12:15:31
15.94	15:56:29
8.78	8:46:50
22.51	22:30:41
17.89	17:53:23
18.10	18:06:07
8.10	8:06:12
20.69	20:41:12
22.29	22:17:07
8.39	8:23:18
15.96	15:57:31
8.03	8:01:38
14.64	14:38:40
6.02	6:01:21
11.60	11:36:03
11.12	11:07:23

(Quadro 44 – continuação)

Quantidade de ligações realizadas (formato decimal)	Quantidade de ligações realizadas (formato horário h/mm/ss)
16.63	16:38:01
15.68	15:40:44
21.26	21:15:38
13.58	13:35:00
7.13	7:07:42
15.39	15:23:32
6.91	6:54:26
8.65	8:38:47
10.98	10:58:59
20.80	20:48:11
17.79	17:47:06
5.25	5:14:48
13.21	13:12:30
12.94	12:56:38
11.43	11:25:39
11.93	11:56:00
12.83	12:49:56
17.80	17:47:58
11.30	11:18:15
9.15	9:09:06
2.66	2:39:20
11.28	11:17:05
9.79	9:47:13
9.65	9:39:11
8.97	8:57:59
9.90	9:53:48
16.10	16:06:07
16.50	16:29:58
10.08	10:05:06

(Quadro 44 – continuação)

Quantidade de ligações realizadas (formato decimal)	Quantidade de ligações realizadas (formato horário h/mm/ss)
8.89	8:53:27
7.57	7:34:16
22.49	22:29:21
9.36	9:21:51
7.19	7:11:33
18.09	18:05:26
15.11	15:06:47
17.04	17:02:28
9.40	9:23:45
1.41	1:24:27
13.48	13:28:30
22.77	22:46:10
10.64	10:38:29
22.39	22:23:34
6.56	6:33:19
4.78	4:46:59
17.89	17:53:37
15.30	15:17:51
18.91	18:54:34
2.46	2:27:24
8.50	8:30:10
7.47	7:28:01
14.09	14:05:11
10.49	10:29:26
7.77	7:46:19
7.90	7:53:52
21.94	21:56:23
9.41	9:24:49
8.20	8:12:03

(Quadro 44 – continuação)

Quantidade de ligações realizadas (formato decimal)	Quantidade de ligações realizadas (formato horário h/mm/ss)
12.64	12:38:19
13.84	13:50:35
2.15	2:09:09
12.62	12:37:27
6.50	6:29:57
6.41	6:24:21
15.74	15:44:32
8.32	8:19:17
7.98	7:58:37
11.97	11:58:06
7.38	7:23:00
13.31	13:18:33
7.67	7:40:17
18.49	18:29:40
10.89	10:53:08
13.72	13:43:14
5.40	5:24:10
3.92	3:55:15
7.91	7:54:34
7.63	7:37:47
12.11	12:06:34
6.19	6:11:32
7.41	7:24:26
8.89	8:53:30
9.38	9:22:39
10.29	10:17:33
17.44	17:26:18
15.21	15:12:25
16.17	16:10:27

(Quadro 44 – continuação)

Quantidade de ligações realizadas (formato decimal)	Quantidade de ligações realizadas (formato horário h/mm/ss)
8.30	8:18:07
16.36	16:21:22
7.00	7:00:09
0.98	0:59:02
14.55	14:33:03
6.03	6:01:59
9.68	9:40:52
6.61	6:36:23
8.74	8:44:40
21.53	21:31:42
8.54	8:32:16
20.23	20:13:59
6.45	6:26:53
18.68	18:40:32
15.33	15:19:57
16.16	16:09:46
14.54	14:32:42
9.93	9:55:56
12.94	12:56:40
14.15	14:08:51
14.23	14:13:44
7.38	7:23:04
8.90	8:54:10
12.18	12:11:04
15.09	15:05:20
12.97	12:58:05
11.78	11:46:32
6.89	6:53:20
12.11	12:06:28

(Quadro 44 – continuação)

Quantidade de ligações realizadas (formato decimal)	Quantidade de ligações realizadas (formato horário h/mm/ss)
8.22	8:13:30
2.48	2:29:02
19.81	19:48:44
13.66	13:39:22
10.95	10:57:05
8.99	8:59:13
9.02	9:01:00
9.02	9:01:03
16.01	16:00:54
9.86	9:51:19
16.09	16:05:29
14.82	14:49:21
8.31	8:18:42
8.26	8:15:49
15.17	15:09:55
15.82	15:49:01
8.95	8:56:55
14.48	14:29:02
19.05	19:02:45
11.23	11:13:42
17.00	17:00:12
2.97	2:57:57
8.26	8:15:45
12.63	12:37:53
9.02	9:01:03
17.36	17:21:45
11.81	11:48:22
17.95	17:56:56
6.60	6:35:56

(Quadro 44 – continuação)

Quantidade de ligações realizadas (formato decimal)	Quantidade de ligações realizadas (formato horário h/mm/ss)
18.39	18:23:29
7.88	7:52:49
12.17	12:10:28
23.33	23:19:32
15.78	15:46:41
8.76	8:45:37
6.69	6:41:34
11.59	11:35:27
15.68	15:40:58
16.60	16:36:03
15.98	15:58:55
7.10	7:06:06
15.50	15:30:01
8.38	8:22:58
18.48	18:28:50
20.06	20:03:53
13.34	13:20:37
16.10	16:06:05
19.03	19:01:46
7.04	7:02:18
15.41	15:24:44
1.04	1:02:07
11.62	11:37:17
10.45	10:27:09
13.17	13:10:18
9.42	9:24:59
2.67	2:40:07
23.88	23:52:47
16.78	16:46:47

Quadro 44 - Ligações sorteadas segundo função degrau

APÊNDICE VI

O Quadro 45 apresenta as ligações sorteadas segundo uma dupla gaussiana.

Quantidade de ligações realizadas (formato decimal)	Quantidade de ligações realizadas (formato horário h/mm/ss)
17.04	17:02:25
12.42	12:24:57
10.38	10:22:38
9.73	9:43:44
16.75	16:45:10
7.31	7:18:23
9.59	9:35:37
11.72	11:42:55
12.64	12:38:39
5.08	5:04:52
6.78	6:46:40
5.22	5:13:08
18.06	18:03:32
1.48	1:29:05
10.00	10:00:07
7.92	7:55:00
11.58	11:35:04
10.81	10:48:35
17.05	17:02:52
13.93	13:55:46
11.44	11:26:10
3.57	3:34:13
7.65	7:38:47
5.86	5:51:25
13.84	13:50:07
14.23	14:13:52
12.73	12:43:34
9.22	9:13:10
15.33	15:19:40
16.95	16:57:04
8.44	8:26:32
10.67	10:40:09
16.46	16:27:30
7.13	7:07:39
10.22	10:13:21
15.60	15:36:11

(Quadro 45 – continuação)

Quantidade de ligações realizadas (formato decimal)	Quantidade de ligações realizadas (formato horário h/mm/ss)
8.59	8:35:32
12.16	12:09:48
14.36	14:21:30
12.37	12:22:20
3.70	3:41:47
16.08	16:04:35
14.14	14:08:35
4.88	4:53:05
7.18	7:10:48
10.84	10:50:14
4.45	4:27:06
12.60	12:36:15
6.20	6:11:44
11.21	11:12:32
5.59	5:35:08
12.19	12:11:28
7.38	7:22:34
2.48	2:28:41
12.26	12:15:43
9.17	9:10:29
20.43	20:25:39
7.05	7:02:51
1.59	1:35:36
13.92	13:55:19
13.51	13:30:37
16.45	16:26:48
12.05	12:03:03
15.29	15:17:10
10.40	10:24:10
12.90	12:53:52
9.88	9:52:39
16.02	16:00:55
14.53	14:31:39
6.49	6:29:28
5.86	5:51:44
13.34	13:20:34
11.05	11:03:06
8.52	8:31:14
7.43	7:26:00
7.18	7:10:47

(Quadro 45 – continuação)

Quantidade de ligações realizadas (formato decimal)	Quantidade de ligações realizadas (formato horário h/mm/ss)
12.35	12:21:01
7.17	7:10:04
6.96	6:57:29
6.65	6:38:48
7.80	7:48:02
8.07	8:04:09
8.51	8:30:31
11.20	11:12:09
19.25	19:14:44
6.67	6:40:28
14.04	14:02:17
17.67	17:40:16
13.42	13:25:02
11.48	11:29:05
7.88	7:52:48
5.27	5:16:23
7.38	7:22:32
14.20	14:12:07
3.34	3:20:14
4.63	4:37:54
10.44	10:26:10
6.53	6:31:55
6.32	6:19:02
12.31	12:18:53
20.54	20:32:37
9.18	9:10:57
9.59	9:35:17
12.75	12:44:55
14.43	14:25:50
5.65	5:39:05
5.78	5:46:55
7.12	7:07:06
9.53	9:31:44
10.10	10:05:56
5.90	5:54:01
11.16	11:09:36
6.87	6:52:30
5.96	5:57:27
6.77	6:45:56

(Quadro 45 – continuação)

Quantidade de ligações realizadas (formato decimal)	Quantidade de ligações realizadas (formato horário h/mm/ss)
15.74	15:44:40
13.46	13:27:26
8.02	8:01:26
7.30	7:17:57
7.19	7:11:36
8.75	8:44:56
12.41	12:24:46
14.29	14:17:24
10.79	10:47:12
9.55	9:33:07
14.75	14:44:57
4.18	4:10:34
14.13	14:08:01
14.49	14:29:21
10.95	10:56:45
10.81	10:48:52
10.52	10:31:24
11.69	11:41:26
18.18	18:10:34
17.33	17:19:47
7.70	7:42:09
7.01	7:00:41
20.74	20:44:42
9.50	9:30:14
8.32	8:19:27
3.72	3:43:29
9.64	9:38:37
8.44	8:26:24
8.67	8:40:15
14.90	14:53:54
5.45	5:26:42
6.53	6:32:01
11.59	11:35:16
6.95	6:57:10
15.13	15:07:48
19.24	19:14:37
12.56	12:33:38
10.18	10:10:53
6.45	6:27:05
10.61	10:36:38
4.53	4:31:43

(Quadro 45 – continuação)

Quantidade de ligações realizadas (formato decimal)	Quantidade de ligações realizadas (formato horário h/mm/ss)
10.10	10:06:06
14.37	14:22:29
6.66	6:39:41
6.02	6:01:15
5.99	5:59:35
6.04	6:02:20
15.64	15:38:32
15.79	15:47:18
9.88	9:52:54
13.92	13:54:56
13.04	13:02:41
11.73	11:43:37
15.05	15:03:10
13.61	13:36:37
5.29	5:17:07
14.72	14:42:58
22.96	22:57:36
4.90	4:54:05
10.28	10:16:39
17.74	17:44:39
11.65	11:38:53
16.79	16:47:09
11.44	11:26:10
16.14	16:08:15
6.04	6:02:06
11.11	11:06:35
14.62	14:37:29
11.74	11:44:32
15.11	15:06:23
8.28	8:16:46
17.08	17:04:53
18.94	18:56:39
11.40	11:23:57
15.57	15:33:54
11.29	11:17:16
8.39	8:23:32
8.32	8:19:29
9.11	9:06:47
6.97	6:58:28

(Quadro 45 – continuação)

Quantidade de ligações realizadas (formato decimal)	Quantidade de ligações realizadas (formato horário h/mm/ss)
4.56	4:33:24
8.61	8:36:36
14.18	14:10:55
11.94	11:56:30
18.83	18:49:39

Quadro 45 - Ligações sorteadas segundo uma dupla gaussiana