MODELOS DE PREVISÃO DE MOBILIDADE HUMANA USANDO DADOS DE FONTES HETEROGÊNEAS

LUCAS MAIA SILVEIRA

MODELOS DE PREVISÃO DE MOBILIDADE HUMANA USANDO DADOS DE FONTES HETEROGÊNEAS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Jussara Marques de Almeida
Coorientador: Humberto Torres Marques Neto

Belo Horizonte
Julho de 2015

Ficha catalográfica elaborada pela Biblioteca do ICEx - UFMG

Silveira, Lucas Maia.

S587m Modelos de previsão de mobilidade humana usando dados de fontes heterogêneas / Lucas Maia Silveira. — Belo Horizonte, 2015.
xviii, 95 f. : il. ; 29cm.

Dissertação (Mestrado) - Universidade Federal de Minas Gerais Departamento de Ciência da Computação.

Orientadora: Jussara Marques de Almeida Gonçalves
Coorientador: Humberto Torres Marques Neto.

1. Computação - Teses. 2. Mobilidade humana 3. Sistemas de comunicação móvel - Teses. 4. Sistema de informação móvel – Teses. I. Orientadora. II. Coorientador. III. Título.

519.6*75 (043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Modelos de previsão de mobilidade humana usando dados de fontes heterogêneas

LUCAS MAIA SILVEIRA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROFA. JUSSARA MARQUES DE ALMEIDA GONÇALVES - Orientadora
Departamento de Ciência da Computação - UFMG

PROF. HUMBERTO TORRES MARQUES NETO - Coorientador
Instituto de Ciências Exatas e Informática - PUC/MG

PROF. ANTONIO ALFREDO FERREIRA LOUREIRO
Departamento de Ciência da Computação - UFMG

PROF. CLODOVEU AUGUSTO DAVIS JÚNIOR
Departamento de Ciência da Computação - UFMG

PROF. EDUARDO COELHO CERQUEIRA
Instituto de Tecnologia - UFPA

Belo Horizonte, 29 de julho de 2015.

Resumo

Entender a mobilidade de grandes grupos de pessoas pode ajudar no planejamento urbano, contenção de doenças ou até em planos de contingência em desastres. Dados de telefonia móvel e de aplicativos georreferenciados são fontes valiosas que permitem analisar a mobilidade de grandes grupos. Nesse sentido, a literatura possui diversos modelos que visam descrever ou prever os padrões de mobilidade humana em uma determinada região em um período de tempo. A maioria desses modelos foi avaliada utilizando dados de uma única fonte, por exemplo, dados de chamadas de telefonia móvel ou dados de GPS obtidos a partir de aplicações Web georreferenciadas. Portanto, a robustez destes modelos a diferentes tipos de dados e, sobretudo, à combinação de dados de múltiplas fontes (dados heterogêneos), é ainda desconhecida.

Nesse contexto, esta dissertação propõe dois modelos de previsão de mobilidade humana que foram projetados para explorar tanto dados de telefonia móvel quanto dados de aplicativos georreferenciados (isolada e conjuntamente). O primeiro modelo, chamado MobDatU, busca prever a mobilidade de uma pessoa em uma área alvo em uma dada janela de tempo baseado na popularidade de cada região da área alvo e nas probabilidades de transição das pessoas entre regiões distintas. Já o segundo modelo, o MobDatU-Contact, busca prever a mobilidade de uma pessoa considerando também a relação de contatos entre as pessoas. Os dois novos modelos foram avaliados e comparados com duas soluções consideradas estado-da-arte, a saber SMOOTH e Leap Graph, em diversos cenários construídos com dados homogêneos e heterogêneos. Os experimentos indicam que o MobDatU atinge resultados melhores ou pelo menos comparáveis ao melhor modelo concorrente em todos os cenários, diferentemente dos modelos alternativos, cujo desempenho é bem mais sensível ao tipo de dado utilizado. Mais ainda, o MobDatU-Contact produziu resultados superiores ao do MobDatU em todos os cenários avaliados, mostrando que a localização dos contatos de uma pessoa pode ser útil na previsão.

Palavras-chave: Mobilidade humana, telefonia móvel, aplicativos georreferenciados.

Abstract

Understanding the mobility of large groups of people can help in urban planning, containment of diseases or even in contigency plans for disaster scenarios. Mobile data and georeferenced applications are valuable sources for examining the mobility of large groups. In this sense, the literature has several models that seek to describe or predict the patterns of human mobility in a particular region over a period of time. Most of these models were assessed using a single data source, e.g. data from mobile phone calls or from georeferenced applications. Therefore, the robustness of these models to different data types, and especially the combination of data from multiple sources (heterogeneous data), is still unknown.

In this context, this dissertation proposes two models to predict human mobility that were designed to explore both mobile data and georeferenced data applications (in an isolated or combined way). The first model, called MobDatU, seeks to predict the mobility of a person in a target area and in a given time window based on the popularity of each region of the target area and the probabilty of transition of the people between two different regions. The second model, the MobDatU-Contact, seeks to predict the mobility of a person considering the contact relationship between people. The two new models as wells as two state-of-the-art models, namely SMOOTH and Leap Graph, were evaluated considering various scenarios with single data source and with multiple data sources. The experiments indicate that the MobDatU always produces results that are better than or at least comparable to the best baseline in all scenarios, unlike the previous models whose performance is much more sensitive to the type of data used. Moreover, the MobDatU-Contact has produced better results than the MobDatU in all evaluated scenarios, showing that the location of the contacts of a person can be useful to predict the human mobility.

Keywords: Human mobility, mobile technology, georeferenced applications.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

1.1 Motivação

O estudo dos padrões de mobilidade humana pode auxiliar no projeto de soluções para melhoria da qualidade dos serviços que suportam a vida urbana [González et al., 2008]. O conhecimento dos padrões de mobilidade humana típicos de uma área alvo, como por exemplo de uma região de um grande centro urbano, pode direcionar o planejamento do fluxo de veículos nas vias de transporte para minimizar a chance de congestionamento durante os horários de trânsito intenso. Além disso, conhecer os padrões de movimentação das pessoas pode possibilitar um melhor planejamento da infraestrutura das redes de telefonia móvel e, consequentemente, uma melhoria da qualidade do serviço oferecido pelas operadoras. Por fim, esses padrões de mobilidade humana em uma determinada região também podem contribuir com a eficácia e eficiência dos sistemas de recomendação baseados na localização dos usuários [Scellato et al., 2011; Noulas et al., 2012b].

Os padrões de mobilidade de uma população em uma área alvo podem ser inferidos a partir da análise de diferentes fontes de dados. Os critérios para classificar diferentes fontes de dados consistem da forma em que uma pessoa é identificada e localizada em uma área alvo. Por exemplo, fontes de dados que contêm chamadas de telefonia móvel. Sabendo que cada chamada que uma pessoa realiza ou recebe está associada a uma antena cuja localização é conhecida, é possível identificar a localização aproximada das pessoas enquanto ele realiza a chamada. Outras fontes de dados que se tornam cada vez mais importantes são aplicativos georreferenciados tais como Foursquare e Twitter[1]. Nessas aplicações, os usuários interagem entre si postando *tweets* ou

[1]Existem *tweets* georreferenciados e também aqueles que não são georreferenciados.

realizando *check-ins*. A localização das pessoas no momento da postagem, capturada pelo dispositivo de GPS (*Global Positioning System*) disponível do dispositivo móvel fica associada a cada *tweet* e *check-in*.

Essas fontes de dados com a localização das pessoas em diferentes horários do dia permitem a construção de modelos analíticos que exploram a distribuição estatística de variáveis relacionadas (*e.g.* distâncias percorridas pelas pessoas) ou mesmo modelos de simulação [Bagrow et al., 2011a; Song et al., 2010], os quais são capazes de prever a mobilidade humana na área geográfica analisada. Estudos anteriores também mostraram que os laços sociais de uma pessoa podem influenciar em sua movimentação. Esses laços sociais podem ser retirados dos dados e também podem ser utilizados para prever a mobilidade humana.

O foco desta dissertação é a formulação de modelos para a previsão de mobilidade humana. Existem diferentes tarefas de previsão de mobilidade humana [Bui et al., 2014]. Uma delas busca prever a mobilidade de um grupo de pessoas [Rhee et al., 2008], visando assim prover subsídios para, por exemplo, o planejamento de vias. Outra tarefa, complementar a anterior, é prever onde uma pessoa estará em um determinado período de tempo futuro [Lee et al., 2012], e assim, inferir a movimentação de cada pessoa. Modelos de previsão que abordam essas duas tarefas estão disponíveis na literatura [Noulas & Mascolo, 2013; Munjal et al., 2011; Dong et al., 2013]. Cada modelo foi avaliado usando um tipo de dado específico. Entretanto, não é conhecido o desempenho desses modelos para fontes de dados distintas (usadas de forma conjunta ou individual). Sendo assim, modelos que apresentam um bom resultado de previsão para uma fonte de dados podem não apresentar o mesmo desempenho com outro tipo de fonte [Noulas & Mascolo, 2013; Scellato et al., 2011].

Além disso, trabalhos que utilizam dados de redes sociais online como fonte de dados mostraram que os contatos entre pessoas podem influenciar na sua movimentação [Cho et al., 2011; Li & Chen, 2009b; Allamanis et al., 2012]. Ou seja, uma pessoa pode ser influenciada por quem ela conhece para ir a um determinado local. Nesse caso, modelos de previsão de mobilidade humana que considerem relações que evidenciam algum conhecimento prévio (*e.g* relações de amizade) poderiam obter resultados melhores quando comparados a um modelo que não utilize tal informação.

1.2 Objetivos

Esta dissertação tem por objetivo principal apresentar modelos que permitam prever onde uma pessoa estará em um instante futuro, considerando uma região alvo e

dados históricos dos padrões de mobilidade naquela região. Busca-se desenvolver modelos eficientes que possam ser parametrizados com dados de diferentes fontes, particularmente dados de telefonia móvel e dados de aplicações georreferenciadas, isolada ou conjuntamente. Busca-se ainda investigar até que ponto pode-se melhorar as previsões utilizando também informações a respeito dos relacionamentos (padrões de contato) entre pessoas. Os modelos desenvolvidos serão amplamente avaliados considerando diferentes cenários de interesse e comparados com soluções existentes consideradas estado-da-arte.

1.3 Contribuições

As contribuições dessa dissertação são:

- A proposição de um novo modelo de previsão de mobilidade humana, MobDatU, que foi projetado para explorar dados de múltiplas fontes, especificamente dados de telefonia móvel e dados de aplicações georreferenciadas. O MobDatU utiliza alguns princípios abordados em dois outros modelos considerados estado-da-arte, o Leap Graph [Dong et al., 2013] e o SMOOTH [Munjal et al., 2011].
- A proposição de um novo modelo de previsão de mobilidade humana que também explora os contatos entre pessoas, o MobDatU-Contact. Este modelo estende o MobDatU ao considerar a probabilidade de encontrar uma pessoa em um local onde pelo menos um de seus contatos se encontre na mesma região que a pessoa.
- Avaliação e comparação dos novos modelos com dois modelos do estado-da-arte, SMOOTH e Leap Graph, em vários cenários.

1.4 Organização

O restante da dissertação está organizado como segue. O Capítulo 2 discute a literatura sobre estudos e modelos de mobilidade humana. O Capítulo 3 apresenta formalmente a tarefa de previsão de mobilidade humana abordada nessa dissertação, detalha o funcionamento dos dois modelos do estado-da-arte considerados na avaliação feita e apresenta os dois novos modelos, o MobDatU e o MobDatU-Contact. A metodologia de avaliação adotada, incluindo as coleções de dados utilizadas nos experimentos é descrita no Capítulo 4, enquanto os resultados são discutidos no Capítulo 5. Conclusões e trabalhos futuros são apresentados no Capítulo 6.

Capítulo 2

Trabalhos Relacionados

Este Capítulo discute os trabalhos relacionados ao tema da dissertação. A Seção 2.1 discute trabalhos que estudaram padrões de mobilidade humana. A Seção 2.2 apresenta trabalhos sobre modelos de previsão de mobilidade humana. Por último, a Seção 2.3 apresenta considerações finais que diferenciam o trabalho desenvolvido nessa dissertação dos estudos anteriores.

2.1 Estudos Sobre Mobilidade Humana

A literatura possui vários estudos sobre a mobilidade humana que se devem ao fato do crescimento de redes de telefonia móvel e do surgimento de aplicações georreferenciadas na sociedade moderna [Vespignani, 2009]. Os estudos sobre os padrões de mobilidade humana podem ajudar em diversas áreas dos setores público e privado. Um exemplo disso é o trabalho de González et al. [2008] que, através de dados de telefonia móvel, identifica os padrões de mobilidade das pessoas em relação ao seu local de trabalho e a sua casa. Essas observações podem contribuir para conduzir o planejamento urbano de cidades. Seja para melhorar o tráfego de veículos, para apoiar a tomada de decisões e ações em situações de emergência e até mesmo planejar melhor a capacidade do sistema de comunicação de uma área alvo [Koo & Chung, 2010; Urgaonkar & Neely, 2011].

Os estudos sobre a mobilidade humana podem ser realizados com diferentes tipos de dados. Entre eles, se destacam dados providos por operadoras de telefonia celular. Essas companhias costumam manter registros de chamadas de todos os seus usuários, incluindo a localização aproximada (com base em suas antenas) nas quais cada chamada começou e terminou, podendo assim identificar quais locais as pessoas visitaram [Balcan et al., 2009]. Outro tipo de dado utilizado nesses estudos são aqueles providos por

aplicações georreferenciadas, como por exemplo o Foursquare e o Twitter. A partir das postagens dessas aplicações é possível extrair a posição de seus usuários [Li & Chen, 2009a].

A partir desses diferentes tipos de dados, análises para identificar e compreender os padrões de mobilidade das pessoas vêm sendo realizadas [Jiang et al., 2012]. Em [Candia et al., 2008] é mostrado que os padrões de mobilidade de uma pessoa variam durante o dia. Ou seja, dependendo da hora do dia uma pessoa pode apresentar diferentes comportamentos em sua movimentação [Silva et al., 2013]. Além disso, outros fatores podem influenciar a movimentação de uma pessoa. Cheng et al. [2011] discutem que questões socioeconômicas podem influenciar os padrões de mobilidade. Em [Cheng et al., 2011] percebe-se que pessoas que moram em grandes centros urbanos tendem a se mover mais do que pessoas que moram em pequenas cidades. Isso acontece por causa da facilidade de locomoção e um maior número de locais que é possível de se visitar nas grandes cidades. [Chiu et al., 2009] mostram que questões geográficas também influenciam na movimentação das pessoas. Ou seja, a movimentação de pessoas entre dois lugares pode ser facilitada ou dificultada devido à existência de barreiras naturais (*e.g.* rios e serras).

Os laços de amizade entre pessoas também influenciam na mobilidade humana [Wang et al., 2011]. [Herrmann, 2003] mostra que existe uma correlação entre as amizades de uma pessoa e de como ela irá se locomover em uma determinada região. Em [Wang & Song, 2015] a relação de amizades de uma pessoa é levada em consideração para entender os seus padrões de mobilidade. Wang & Song [2015] mostraram que além da relação de amizade influenciar na mobilidade humana, quanto mais perto a localização de um amigo está de outro, maiores são as chances de eles se encontrarem. Ou seja, assim como mostrado em [Becker et al., 2013], usuários tendem a se locomover mais para localidades próximas de sua posição atual do que para locais mais distantes.

Esses laços de amizades utilizados para entender os padrões de mobilidade, podem ser inferidos a partir de diferentes maneiras e para os diferentes tipos de dados. Em [Eagle et al., 2009] é inferido os laços de amizade a partir do número de chamadas que duas pessoas realizam entre si e também considera a proximidade entre os localis que ambas visitam. Já em [Li & Chen, 2009b], por utilizar dados de aplicativos georreferenciados (coletados da rede social Brightkite), foi possível retirar diretamente a relação de amizade das coleções. Ou seja, os dados do Brightkite possuem como parâmetro a lista de amigos de cada pessoa. Tanto inferindo, quanto retirando as relações de amizades foi possível perceber que a mobilidade humana sofre influência dos laços sociais [Cho et al., 2011].

Além dos fatores que influenciam na movimentação diária das pessoas, os padrões de mobilidade humana podem sofrer alterações repentinas. Por exemplo, quando ocorre algum tipo de emergência, o padrão de atividade humana se altera, gerando um novo padrão de mobilidade. Além do fato que há um aumento quase instantâneo na utilização da rede móvel nas proximidades do local do evento [Bagrow et al., 2011b]. Com essa grande quantidade de dados que inundam a rede repentinamente e permanecem altos por um período de tempo considerável é possível identificar e classificar aquele evento como uma emergência. Essas emergências podem ser desde desastres naturais (*e.g.* furacões), até manifestações.

Wang et al. [2009] argumentam que a variação do volume de dados pode indicar o tipo de evento que pessoas estão experimentando, por exemplo, uma emergência, um evento de esporte, ou um concerto. Além da variação do volume de dados, o comportamento das pessoas pode variar de acordo com o tipo de evento [Calabrese et al., 2010]. Sendo assim, eventos diferentes, por exemplo um show e uma partida de futebol, tendem a apresentar padrões de mobilidade humana distintos em relação como as pessoas chegam e saem do local do evento [Erman & Ramakrishnan, 2013]. Em [Xavier et al., 2012] é analisada essa alteração nos padrões de mobilidade em partidas de futebol. Os autores Xavier et al. [2012] mostram que a quantidade de ligações aumenta gradualmente no início, diminui um pouco durante e retorna a aumentar na saída do local por um período curto, e depois volta para o nível normal de utilização da rede móvel.

Por último, o trabalho de Gonzalez [2013] mostra que os padrões de mobilidade humana tendem a possuir características similares para diferentes localidades. Com isso, se torna possível comparar a movimentação das pessoas em diferentes locais superando as barreiras culturais, nacionais e organizacionais [Noulas et al., 2012a]. Porém, como dito no trabalho de Blondel et al. [2015], alguns padrões da movimentação das pessoas são específicos de cada local. Por exemplo, quanto mais locais próximos da localização da pessoa maior a chance de ela permanecer mais tempo se movendo.

2.2 Previsão de Mobilidade e Modelos

Uma forma de aplicar os estudos dos padrões de mobilidade humana é na criação de modelos para prever a movimentação de pessoas. Esses modelos utilizam diferentes tipos de dados e exploram diversas estratégias para tentar prever com maior precisão a movimentação das pessoas [Bui et al., 2014]. Alguns destes modelos buscam prever as trajetórias das pessoas utilizando dados de GPS [Zheng & Xie, 2011] ou dados

de telefonia móvel [Dong et al., 2013]. Outros modelos realizam a previsão a partir de distribuições estatísticas que capturam padrões específicos identificados nos dados, tais como a distribuição da distância percorrida por um usuário [Lee et al., 2009; Munjal et al., 2011].

Uma das formas de si realizar a previsão de mobilidade humana é a partir de modelos sintéticos [Tasse & Glass, 2008; Navidi et al., 2004]. Esses modelos geram dados que não representam os padrões de mobilidade. Isso pois, não foram observados dados reais para projetá-los. Sendo assim, apesar de apresentarem uma abordagem simples de ser desenvolvida, os modelos sintéticos não conseguem representar os padrões de mobilidade observados de dados reais [Musolesi & Mascolo, 2008]. Isso, acaba afetando a precisão desse tipo de modelo para cenários reais [Walsh et al., 2008].

Os modelos que buscam aprender os padrões de mobilidade humana podem utilizar dados provenientes de diferentes fontes, tais como dados de GPS, de operadoras de telefonia móvel e de aplicações georreferenciadas. Um exemplo, é o modelo proposto por Rhee et al. [2008] que considera as características de movimentação das pessoas. Rhee et al. [2008], ao analisarem seus dados de GPS, perceberam que a movimentação humana segue uma distribuição de cauda pesada. Ou seja, as pessoas percorrem distâncias pequenas com mais frequência do que distâncias longas. Um outro exemplo é o modelo proposto por Lee et al. [2009] que além de considerarem que a movimentação humana segue uma distribuição de cauda pesada, também perceberam que as pessoas tendem a se mover para locais que são mais visitados.

Tostes et al. [2013], além de notarem que as pessoas tendem a visitar lugares mais populares, perceberam que a popularidade de uma localidade pode variar de acordo com o período do dia (madrugada, manhã, tarde e noite). Com isso, em [Tostes et al., 2013] é proposto um modelo para prever o fluxo de veículos nas vias e que considera as variações no tráfego em diferentes períodos do dia. De forma similar, outros trabalhos, como Jeung et al. [2008], exploram diversos padrões observados nos dados, tais como o tipo de lugar que as pessoas visitam (*e.g.* casa, trabalho e lazer), a frequência de visitação e a regularidade dos padrões de locomoção de cada pessoa. Em Noulas et al. [2012b], os autores exploram os locais visitados bem como as distâncias de locomoção de um usuário dentro de uma mesma região para, não só, prever sua localização em um momento futuro, mas também poder recomendar lugares que ele possa achar interessantes.

Em [Munjal et al., 2011] é realizado uma abordagem mais simples, porém que ainda representa padrões reais de mobilidade. Os autores Munjal et al. [2011] perceberam que poderiam prever a posição de uma pessoa a partir do seu padrão de distância percorrida, tempo em que ficou parado e a popularidade da região. Para isso, eles

retiraram dos dados a distribuição das distâncias que a pessoa percorreram em um deslocamento, o tempo entre deslocamentos que ela ficou parada e a popularidade de cada região. Assim, como em [Kim et al., 2006], percebeu-se, a partir dos dados de GPS, que a maioria das pessoas tendem a percorrer distâncias pequenas. Além disso, locais que muitas pessoas visitam tendem a receber mais pessoas do que locais menos populares [Liu et al., 2010].

Em [Noulas & Mascolo, 2013] e [Wang & Song, 2015] foi observado que além das aplicações georreferencias(*e.g.*,Foursquare), os dados de telefonia móvel também seguem a mesma distribuição de movimentação que foi observada em [Lee et al., 2009] e [Munjal et al., 2011]. Além disso, os autores Noulas & Mascolo [2013] consideraram em seu modelo que pessoas tendem a ir com mais frequência a lugares específicos, dependendo do dia da semana. Por exemplo, pessoas tendem a visitarem teatros e cinemas com mais frequência nos fim de semanas do que durante a semana.

Em [Dong et al., 2013] a previsão de mobilidade humana é realizada observando a trajetória das pessoas. Os autores de [Dong et al., 2013] perceberam que as pessoas tendem a seguir uma mesma trajetória durante o dia. Eles criaram um modelo que identifica essas trajetórias a partir de dados de telefonia para prever a movimentação humana. Já Isaacman et al. [2012], que também utilizam dados de telefonia móvel, propõem um modelo que prevê a localização das pessoas baseado no tipo de lugar que elas visita. Ou seja, eles analisam o número de vezes que cada pessoa se movimenta entre sua casa e trabalho e identificam quais as chances de ela visitar cada local. No caso de [Csáji et al., 2013] e [Song et al., 2010] foi considerado que pessoas tendem a visitar lugares conhecidos com mais frequências do que lugares novos.

Um outro fator que vem sendo utilizado para realizar a previsão de mobilidade humana é a amizade. Os modelos propostos em [Musolesi & Mascolo, 2007] e [Nguyen & Szymanski, 2012] exploram não somente os padrões de distância percorrida e locais visitados, mas também as relações de amizade entre os usuários (inferidas a partir dos dados coletados). Ou seja, assim como em [Davis Jr. et al., 2011], eles inferem que uma pessoa pode estar em um local, se seus amigos também se encontram na mesma localidade.

Diferente de [Musolesi & Mascolo, 2007], que utiliza os próprios laços de amizade que são retirados das coleções, também é possível inferir os laços sociais entre duas pessoas. No modelo de Alharbi & Zhang [2014] é considerado que laços sociais podem ser inferidos a partir dos locais que as pessoas visitam. Sendo assim, pessoas que visitam os mesmos lugares no mesmo intervalo de tempo podem ser consideradas amigas. [Scellato et al., 2011], além de considerar os mesmos locais visitados, também considera uma margem de erro. Ou seja, um laço social existirá se duas pessoas estão

em locais próximos, mas não necessariamente na mesma posição.

Em [Jia et al., 2014] é proposto um modelo que além de considerar os amigos de uma pessoa para prever sua localização, considera que os laços sociais podem alterar de acordo com o tempo. Ou seja, uma pessoa pode possuir amigos diferentes com o passar do tempo. Os autores Jia et al. [2014] consideram que novos laços de amizade podem surgir de acordo com o número de vezes que duas pessoas frequentam um mesmo local. Ideia similar é proposta em [Allamanis et al., 2012] que além de considerar os locais que as pessoas visitam para considerar um laço de amizade, considera os amigos que ambos tem em comum.

Por último, Soper [2012] discute a importância dos modelos de previsão de mobilidade humana no desenvolvimento tecnológico e na melhora da qualidade de vida de uma região. Porém esses modelos também podem violar a privacidade das pessoas e assim fazendo com que os setores público e privado possam controlar esse tipo de informação e utilizá-la para proveito próprio. Sendo assim, apesar dos benefícios das informações providas por um modelo de previsão de mobilidade humana, é necessário criar medidas para que os dados das pessoas também sejam preservados.

2.3 Considerações Finais

A previsão da mobilidade humana pode auxiliar na prevenção de doenças e desastres bem como no planejamento urbano. Por exemplo, os trabalhos de González et al. [2008] e Balcan et al. [2009] mostram que o conhecimento sobre a mobilidade das pessoas em uma determinada região pode auxiliar na tomada de decisões mais rápidas para evitar a disseminação de uma doença, amenizar os danos causados por desastres e até mesmo evitar engarrafamentos durante as horas de pico. Além disso, tal conhecimento também pode auxiliar as operadoras a aperfeiçoarem seus serviços de telefonia móvel [Xavier et al., 2012].

Entretanto, os modelos de mobilidade humana disponíveis na literatura foram propostos ou avaliados considerando apenas uma fonte única de dados, sejam chamadas de telefonia móvel [Dong et al., 2013; Csáji et al., 2013; Balcan et al., 2009] sejam dados de GPS coletados de aplicações georreferenciadas [Lee et al., 2009; Rhee et al., 2008; Munjal et al., 2011; Musolesi & Mascolo, 2007; Noulas et al., 2012a; Nguyen & Szymanski, 2012; Cho et al., 2011]. A robustez desses modelos a dados de fontes diversas, seja isolada ou conjuntamente, é o foco deste trabalho. Para tanto são selecionados dois modelos de referência, o SMOOTH [Munjal et al., 2011] e o Leap Graph [Dong et al., 2013]: enquanto o primeiro foi avaliado anteriormente somente a partir

de dados de GPS, o segundo foi avaliado para dados de telefonia móvel. Mais ainda, é proposto um novo modelo, chamado MobDatU, que herda princípios dos dois modelos de referência mas é projetado para explorar dados de fontes heterogêneas. Além disso, é proposto o MobDatU-Contact, que utiliza os mesmos princípios do MobDatU e também considera os laços sociais entre as pessoas para realizar a previsão de mobilidade.

Capítulo 3

Previsão de Mobilidade

Este capítulo apresenta na Seção 3.1 a tarefa de previsão de mobilidade humana abordada nesta dissertação. A seguir, é descrito o funcionamento dos dois modelos de referência, SMOOTH e Leap Graph (Seção 3.2). Por fim, são introduzidos na Seção 3.3 os modelos de previsão propostos, o MobDatU e o MobDatU-Contact. A Tabela 3.1 apresenta a notação usada neste capítulo.

Tabela 3.1: Notação - Previsão de Mobilidade

Variável	Definição
U	conjunto de usuários
R	conjunto de regiões
T	conjunto de janelas de tempo
$\mathcal{D}^{treino}$	conjunto de treino composto por tuplas $< u_i, r_i, t_i >$, onde $u_i \in U$, $r_i \in R$, e $t_i \in T$, utilizado para aprender os padrões de mobilidade humana e derivar os modelos de previsão
$\mathcal{D}^{teste}$	conjunto de teste composto por tuplas $< u_i, r_i, t_i >$, onde $u_i \in U$, $r_i \in R$, e $t_i \in T$, utilizado para avaliar modelos de previsão de mobilidade humana
p_i	popularidade da região i aprendida em $\mathcal{D}^{treino}$
d	parâmetro que define o tamanho de uma região, podendo ser o tamanho do lado, no caso de regiões retângulares, ou o raio para o caso de regiões circulares
f_{dist}	distribuição das distâncias percorridas em um deslocamento por um usuário em U
f_{pause}	distribuição do tempo de um usuário em U que permaneceu parado entre deslocamentos sucessivos
t_{max}	número máximo de janelas de tempo em $\mathcal{D}^{treino}$
C_i	conjunto de contatos do usuário i aprendidos a partir de $\mathcal{D}^{treino}$
$L_{i,j,t}$	evento que indica que o usuário i está localizado na região r_j na janela de tempo t
$C^{\geq n}_{i,j,t}$	evento que indica que pelo menos n contatos do usuário i estão na região r_j na janela de tempo t
$C^{k}_{i,j,t}$	evento que indica que exatamente k contatos do usuário i estão localizados na região r_j na janela de tempo t

3.1 Definição do Problema

A tarefa de previsão tratada nesta dissertação pode ser definida como segue. Sejam uma área alvo R, consistindo de um conjunto de regiões $r_i \in R$, um conjunto de

usuários $u_i \in U$, um intervalo de tempo $\mathcal{T}$ subdividido em janelas de tempo (*e.g.*, uma hora) e um conjunto t_i de janelas de tempo. Sejam ainda um conjunto de treino $\mathcal{D}^{treino}$ e um conjunto de teste $\mathcal{D}^{teste}$, consistindo de tuplas $< u_i, r_i, t_i >$ que indicam que u_i estava na região r_i em t_i. Deseja-se construir um modelo para prever em que região r_i um dado usuário u_i estará em momento t_i utilizando apenas os dados de treino $\mathcal{D}^{treino}$. Deseja-se ainda avaliar o modelo utilizando os dados de teste, ou seja, utilizar o modelo desenvolvido para prever a localização (*i.e.*, região) r_i para cada tupla $< u_i, ?, t_i >$ no conjunto de teste $\mathcal{D}^{teste}$.

Note que a definição das regiões r_i pode ser feita de diversas maneiras. Por exemplo, cada região pode ser definida por um ponto central x_i, y_i (*e.g.*, as coordenadas de uma antena para o caso de dados de telefonia móvel) e um raio d. Alternativamente, cada região pode representar uma área quadrada com centro x_i, y_i e lado d (área total d^2). A definição das regiões adotada por cada modelo é explicitada nas seções seguintes. Note também que, nesta dissertação, o intervalo de tempo total $\mathcal{T}$ é discretizado em janelas consecutivas t_i para fins de computação dos deslocamentos dos usuários. Em outras palavras, a localização de um usuário é predita considerando a granularidade de tempo de t_i.

A divisão dos dados disponíveis em treino e teste também pode ser feita de diversas formas, dependendo da disponibilidade dos dados. Entretanto, tal divisão deve respeitar restrições temporais, ou seja, os dados de treino devem preceder os dados de teste (*i.e.*, $\forall <*, *, t_i> \in \mathcal{D}^{treino}$ e $\forall <*,*,t_j> \in \mathcal{D}^{teste}$, $t_i < t_j$). Mais ainda, considerando que os padrões de mobilidade humana variam ao longo do dia ou mesmo em dias diferentes (*e.g.*, dias de semana e fins de semana) [Xavier et al., 2012], é desejado que os dados de treino tenham sido obtidos em períodos comparáveis com aqueles do conjunto de teste. Por exemplo, se é desejado prever a localização de usuários entre 8 e 9 da manhã, deve-se utilizar dados coletados do mesmo período em dias anteriores ou dados coletados em um período imediatamente anterior. A definição dos conjuntos de treino e teste em nossos experimentos será discutida na Seção 4.1.

3.2 Modelos de Referência

Esta seção descreve os principais componentes dos dois modelos de referência adotados neste trabalho: o SMOOTH [Munjal et al., 2011] e o Leap Graph [Dong et al., 2013]. Em nossos experimentos, nós utilizamos as implementações do SMOOTH e do Leap Graph disponibilizadas pelos autores em suas respectivas páginas[1].

[1]SMOOTH: https://toilers.mines.edu; Leap Graph: https://www.cs.utexas.edu/∼wdong86/

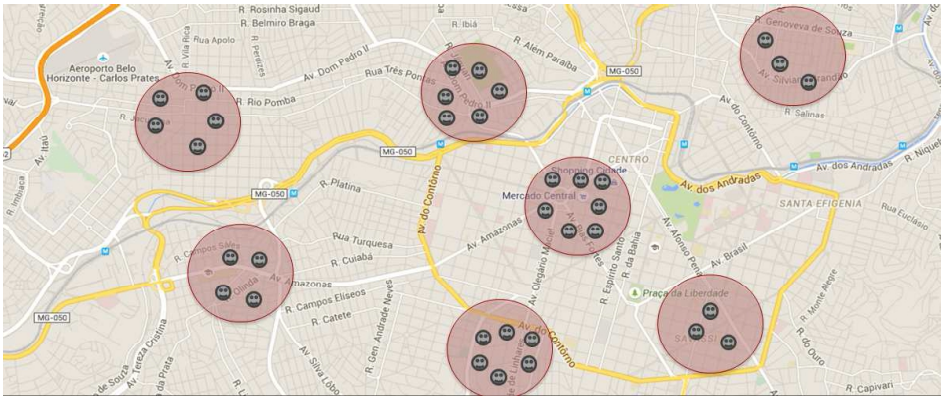


Figura 3.1: SMOOTH - Regiões de Interesse com Pessoas

3.2.1 SMOOTH

O modelo SMOOTH, proposto por Munjal et al. [2011], captura a locomoção de um grupo de usuários U em uma área bidimensional, composta por um conjunto de regiões de interesse. Na Figura 3.1 é possível ver um exemplo dessas regiões, no qual cada região r_i é definida por coordenadas x_i, y_i dentro da área simulada, um raio de distância d e probabilidade p_i de um usuário se deslocar para ela. A probabilidade p_i, extraída do conjunto de treino, captura a popularidade da região r_i, isto é, o número esperado de pessoas que visitam r_i. A Figura 3.1 mostra as regiões com as pessoas que as visitaram.

A ideia básica do SMOOTH é simular a locomoção dos usuários U em uma sequência de passos, onde cada passo representa uma janela de tempo com duração t_i minutos. Em cada passo, o modelo simula a locomoção de cada usuário $u_i \in U$ a partir de duas distribuições específicas extraídas do conjunto de treino: a distribuição das distâncias percorridas f_{dist} e a distribuição dos tempos de pausa f_{pausa}. Como em [Munjal et al., 2011], observamos que tais distribuições seguem leis de potência para todos os cenários simulados. Tais distribuições são caracterizadas por três parâmetros, a saber α, valor mínimo (min) e valor máximo (max), com uma distribuição acumulada complementar dado por $P(X > x) = (\frac{x*max^{\alpha} - x*min^{\alpha} - max^{\alpha}}{(x*max^{\alpha})*(x*min^{\alpha})})^{-\alpha}$[Klebanov, 2003]. Por exemplo as Figuras 3.2 e 3.3 mostram, respectivamente, as distribuições f_{dist} e f_{pausa} para um cenário específico de dados do Twitter para o Rio de Janeiro dia 29/06/2014. Os valores de α, min e max são 1,67, 23 metros, 4002 metros para f_{dist} e 1,74, 4s e 14000s ($2h$) para f_{pausa}.

A simulação ocorre da seguinte maneira. Em cada passo, para cada usuário que não está em pausa (explicado posteriormente), primeiramente é computada a direção de deslocamento em função da sua localização atual e das probabilidades associadas a cada região $r_i \in R$. Em seguida, é selecionada aleatoriamente uma distância de deslocamento

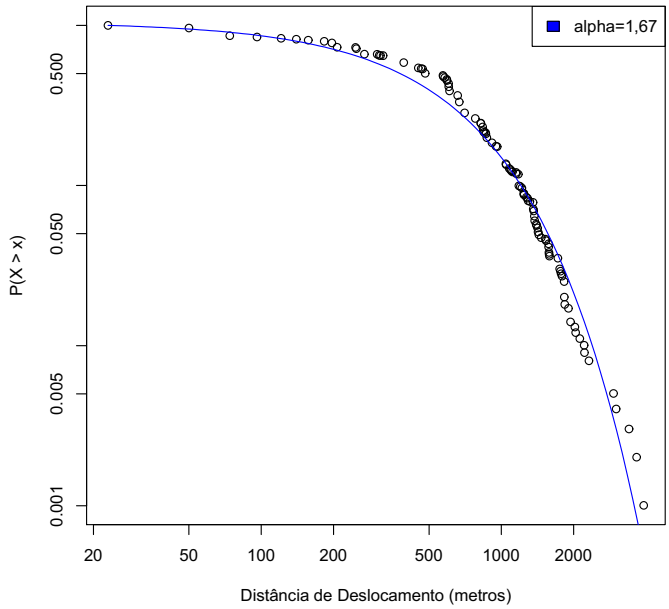


Figura 3.2: Distância dos Deslocamentos para *Tweets* - Rio de Janeiro 29/06/2014

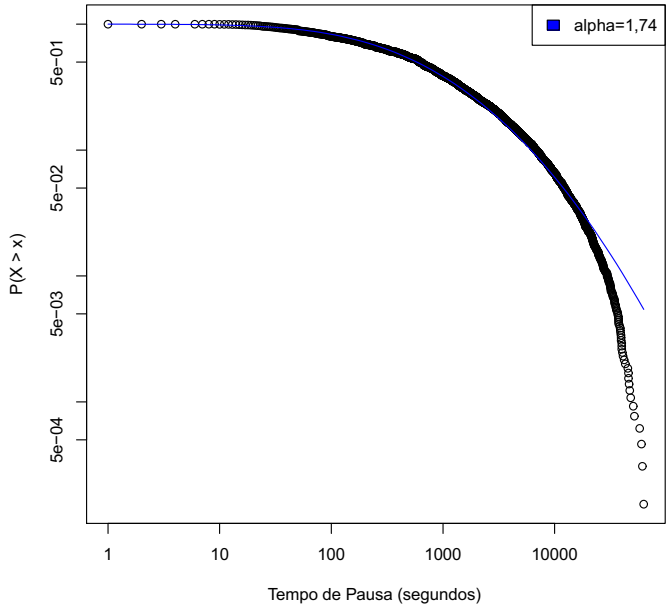


Figura 3.3: Tempo de Pausa entre Deslocamentos para *Tweets* - Rio de Janeiro 29/06/2014

utilizando f_{dist}, e o deslocamento é simulado. Por fim, é escolhido aleatoriamente um tempo de pausa utilizando f_{pausa}. O usuário permanecerá neste local durante o intervalo de tempo f_{pausa} selecionado. Parte desse processo pode ser visualizado na Figura 3.4

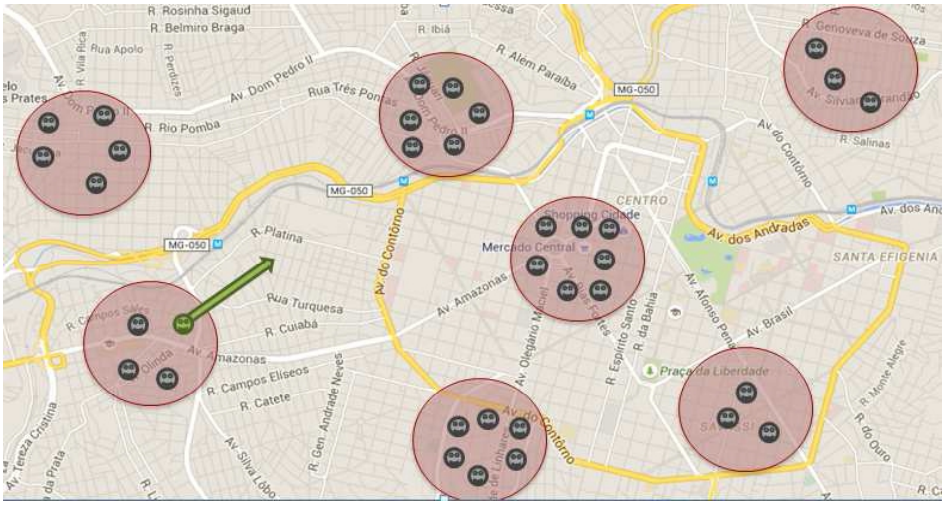


Figura 3.4: SMOOTH - Deslocamento de Usuário

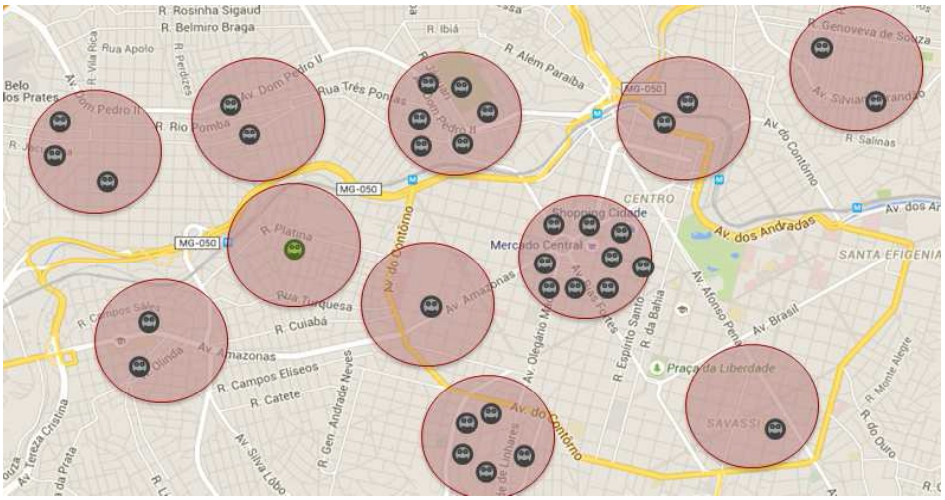


Figura 3.5: SMOOTH - Após Deslocamento do Usuário

para o qual existe um usuário com uma seta que representa qual será a direção e a distância de seu deslocamento para o próximo intervalo de tempo.

Ainda para cada usuário, é verificado se ele está dentro do raio de cobertura d que define uma região previamente identificada, ou seja, se o usuário está no raio de cobertura d, ele é associado àquela região. Caso contrário, é criada uma nova região r_i em R definida pela sua localização atual. Como exemplo, a Figura 3.5 mostra que regiões foram criadas para os usuários, que após deslocar-se, não se encontravam em nenhuma região já existente. Ao final de cada passo, as probabilidades associadas a cada região (inclusive as novas) são recomputadas.

Sendo assim, o treinamento do modelo é feito em duas etapas. A primeira consiste em extrair as distribuições f_{dist} e f_{pausa} bem como as regiões $r_i \in R$ e suas probabilidades p_i do conjunto $\mathcal{D}^{treino}$. Todas as regiões que aparecem no conjunto de treinamento

são inicialmente introduzidas no conjunto R com as probabilidades correspondentes. Na segunda etapa do treinamento, as posições iniciais dos usuários são determinadas a partir das probabilidades associadas às regiões. Além disso, são simulados os deslocamentos dos usuários usando as distribuições $f(dist)$ e $f(pausa)$, por um número de janelas de tempo igual a $\mathcal{D}^{treino}$. Novas regiões descobertas durante esta fase são inseridas no conjunto R, com suas probabilidades.

Durante a fase de teste, os deslocamentos dos usuários do conjunto $\mathcal{D}^{teste}$ são simulados utilizando o modelo aprendido, mantendo o conjunto R fixo e considerando a localização inicial de cada usuário dada pela sua primeira aparição em $\mathcal{D}^{teste}$. As regiões visitadas pelos usuários durante a simulação são comparadas com os demais dados do conjunto $\mathcal{D}^{teste}$ para avaliar a precisão do modelo.

3.2.2 Leap Graph

Em [Dong et al., 2013], os autores investigaram como utilizar dados de telefonia móvel para prever a mobilidade de usuários. Esses dados correspondem a um conjunto de chamadas telefônicas. A cada chamada estão associados um identificador único do usuário, os instantes de início e fim da chamada, bem como as coordenadas (latitude e longitude) das antenas onde a chamada foi iniciada e finalizada, bem como as identificações dos setores utilizados nessas antenas[2].

No modelo proposto por Dong et al. [2013], chamado Leap Graph, cada região $r_i \in R$ corresponde a uma antena, sendo definida pelas suas coordenadas e por um raio d que representa sua área de cobertura (Figura 3.6). Assim, uma chamada do usuário u_i associada à antena correspondente à região r_i em um instante t indica a presença de u_i naquela região na janela de tempo que inclui t. O modelo tenta inferir os deslocamentos de cada usuário a partir de um grafo que captura as trajetórias dos usuários entre as regiões de R.

A fase de treinamento portanto consiste primeiramente em criar um grafo de trajetórias para cada usuário a partir dos dados em $\mathcal{D}^{treino}$. No grafo G_i criado para o usuário u_i, cada vértice corresponde a uma região. Uma aresta entre r_i e r_j é adicionada toda vez que: (i) u_i fez uma chamada que foi iniciada em r_i e finalizada em r_j; ou (ii) u_i fez duas chamadas consecutivas, a primeira em r_i e a segunda em r_j.

Como proposto originalmente, o Leap Graph objetiva prever a próxima região em que um usuário estará dada a sua localização atual. Portanto, ele não considera a dimensão tempo e explora apenas as transições entre antenas feitas por cada usuário.

[2]Cada antena é dividida em 3 setores de 120^o, cada um responsável por cerca de um terço da área de cobertura da antena.

Figura 3.6: Leap Graph - Regiões das Antenas

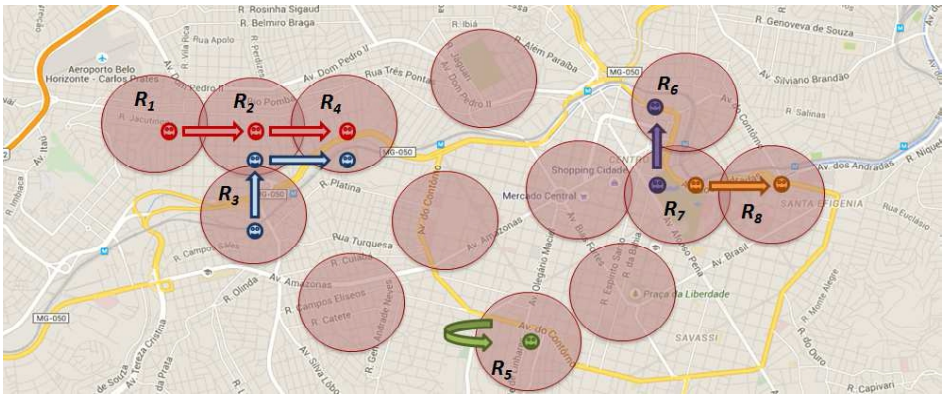


Figura 3.7: Leap Graph - Deslocamento dos Usuários

Para torná-lo comparável ao SMOOTH e aos nossos novos modelos (que consideram o deslocamento ao longo do tempo) e aplicá-lo à tarefa de predição alvo, nós adicionamos laços (arestas da região r_i para ela mesma) para capturar os períodos entre chamadas consecutivas de um mesmo usuário. Durante tais períodos foi assumido que o usuário permaneceu a metade do tempo em uma região e a metade seguinte na outra. Ou seja, dadas duas chamadas consecutivas realizadas pelo mesmo usuário u_i, a primeira na região r_i no instante t_i e a segunda na região r_j no instante t_2, foram adicionados laços de r_i para r_i e de r_j para r_j. A Figura 3.7 mostra três usuários e todas as regiões que se deslocaram durante todas as janelas de tempo do treinamento.

Os grafos criados são então combinados em um grafo único ponderado G que representa os deslocamentos da população de usuários em $\mathcal{D}^{treino}$. Para tal, os grafos de usuários G_i são ordenados pelo instante da primeira chamada de cada usuário em $\mathcal{D}^{treino}$ e processados conforme esta ordem. As arestas de todos os grafos são combinadas em G, sendo que o peso de uma aresta corresponde ao número de grafos de usuários em que ela aparece. Porém, para trajetórias cobrindo n ou mais arestas que aparecem em múltiplos grafos (n é parâmetro do modelo), são considerados apenas a trajetória e os trechos que a sucedem no primeiro grafo processado. Em outras palavras, suponha que

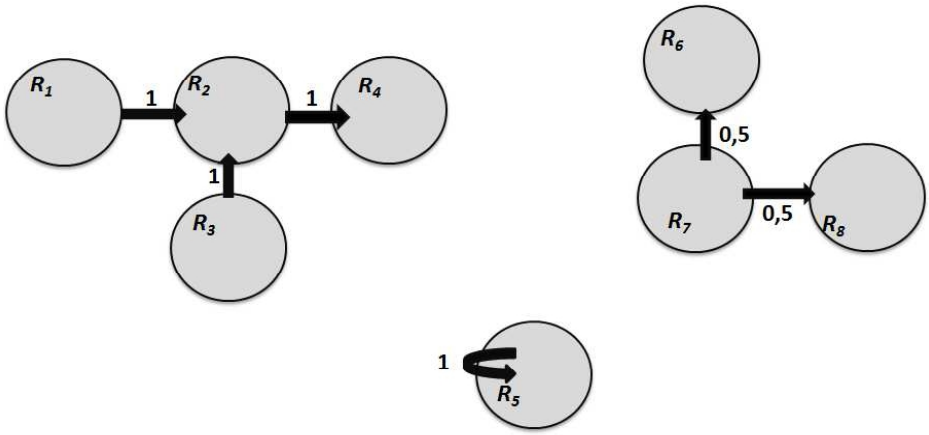


Figura 3.8: Leap Graph - Cadeia de Markov

o grafo G_1 contenha a trajetória $\{r_1, r_2, r_3, r_4\}$ e que grafo G_2 contenha a trajetória $\{r_1, r_2, r_3, r_5\}$. Para $n = 2$, o grafo G conterá as arestas $\{r_1, r_2, r_3, r_4\}$, todas com peso 1, já que a trajetória $\{r_1, r_2, r_3\}$ aparece nos dois grafos. Conforme os autores, essa medida é tomada para se evitar a dupla contabilização das mesmas trajetórias. Ao final da combinação, os pesos das arestas são normalizadas de forma que os pesos das arestas de saída de cada vértice r_i totalizem 1, conforme mostrado na Figura 3.8.

A aplicação do modelo Leap Graph, durante a fase de teste, consiste em simular o grafo G produzido durante o treinamento como uma cadeia de Markov. Na Figura 3.8 é representado o grafo produzido pelos deslocamentos dos usuários que foi mostrado na Figura 3.7. Para cada usuário, a sua posição inicial é extraída da sua primeira chamada no conjunto $\mathcal{D}^{teste}$. Esta corresponde a um estado da cadeia. A cadeia é então simulada para inferir a posição do usuário em sucessivos passos. Em nossos experimentos, utilizamos $n = 2$, pois, foi a escolha que levou aos melhores resultados em [Dong et al., 2013].

3.3 Novos Modelos de Previsão de Mobilidade

Esta seção apresenta os novos modelos de mobilidade propostos - MobDatU e MobDatU-Contact. A Seção 3.3.1 introduz o MobDatU ressaltando as similaridades e principais diferenças entre ele e os modelos de referência. A Seção 3.3.2 aprensenta o MobDatU-Contact, uma extensão do MobDatU que explora os padrões de contato entre as pessoas.

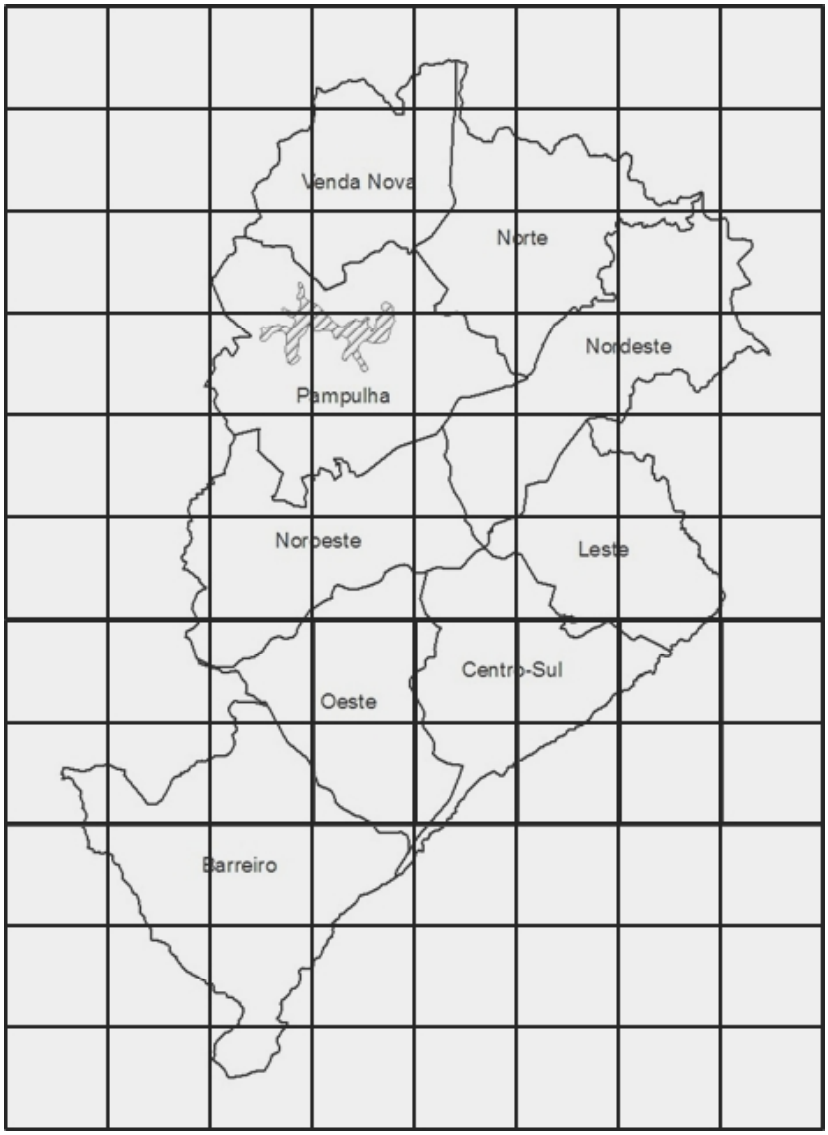


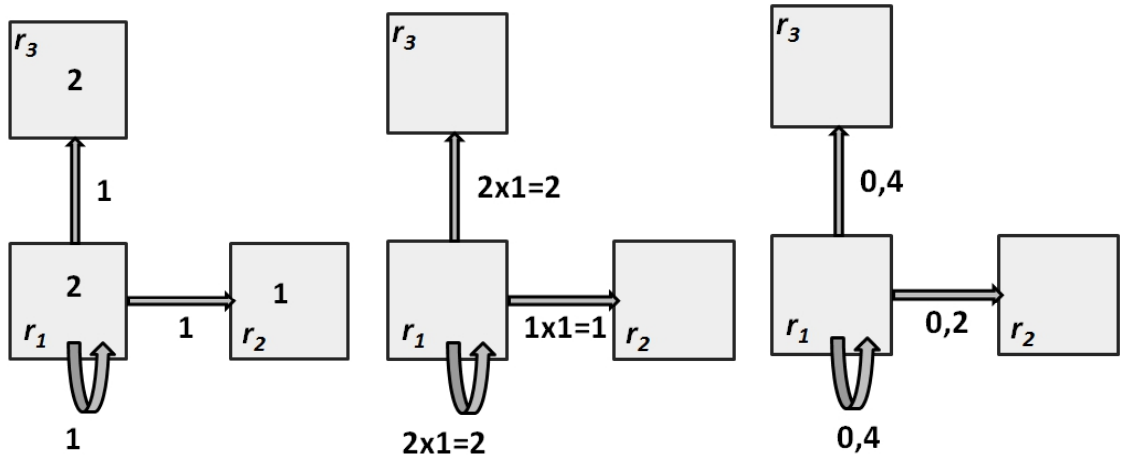
Figura 3.9: MobDatU - Divisão das Regiões como *Grid*

3.3.1 MobDatU

O novo modelo MobDatU tem como objetivo explorar dados de fontes heterogêneas, de forma conjunta ou individual, para capturar a movimentação dos usuários entre as regiões de R. A área total simulada pelo modelo é dividida em regiões quadrangulares não sobrepostas r_i, como mostrado na Figura 3.9. Cada região r_i é definida por um centro x_i, y_i e um lado d[3].

O MobDatU herda alguns aspectos dos dois modelos de referência. Por exemplo, assim como no SMOOTH, a cada região r_i é associada uma medida de popularidade, que, no MobDatU, representa o número de usuários que visitaram r_i no conjunto de

[3]O modelo também foi executado utilizando regiões circulares de raio d, apresentando resultados similares aos reportados nesta dissertação.



(a) Popularidade das Regiões e Frequência de Transição entre Regiões

(b) Definindo Peso das Arestas

(c) Normalização dos Pesos das Arestas

Figura 3.10: Cálculo da Probabilidade de Transição

treino $\mathcal{D}^{treino}$ (podendo ser 0). De forma similar ao Leap Graph (e diferentemente do SMOOTH), o MobDatU simula a movimentação dos usuários entre as regiões a partir de um grafo de transições. Entretanto, diferentemente do Leap Graph, a criação deste grafo não parte das trajetórias individuais de cada usuário e também não inclui o descarte de trajetórias com prefixo comum durante o processo de combinação dos grafos de usuários.

Ao invés disto, o MobDatU cria um grafo de transições onde o peso associado a uma aresta (r_i, r_j) representa o número de pessoas que fizeram a transição entre as duas regiões no conjunto $\mathcal{D}^{treino}$. Assim como no Leap Graph, transições *self loop* são introduzidas para capturar os períodos entre chamadas ou *tweets* sucessivas de um mesmo usuário.

A fase de treinamento do modelo consiste, portanto, em aprender o grafo de transições (incluindo os pesos das arestas) e as popularidades de cada região a partir do processamento do conjunto $\mathcal{D}^{treino}$. Ao final, os pesos de todas as arestas são recomputados para capturar as popularidades de cada região destino. A Figura 3.10 mostra esse processo. Cada quadrado corresponde a uma região, e o número no centro do quadrado indica a popularidade da região correspondente. Por exemplo, as popularidades de r_1 e r_3 são 2, enquanto a popularidade de r_2 é 1 (Figura 3.10(a)).

Primeiramente os pesos de cada aresta (r_i,r_j) é multiplicado pela popularidade da região destino como mostrado na Figura 3.10(b). Em seguida os pesos das arestas são todos normalizados de forma a refletir as probabilidades de transição, ou seja, a soma dos pesos de todas as arestas saindo de um mesmo vértice devem ser igual a 1.

Isso é mostrado na Figura 3.10(c).

Esse é um ponto em que o MobDatU difere tanto do Leap Graph, que considera somente as probabilidades de transição, quanto do SMOOTH, que explora somente as popularidades de cada região. O MobDatU considera que ambos aspectos podem influenciar a trajetória de um usuário: se por um lado os usuários tendem a visitar locais específicos dependendo da sua localização atual (como mostrado em [Dong et al., 2013]), por outro, a popularidade de uma região também influencia a movimentação dos usuários [Munjal et al., 2011; Noulas et al., 2012a].

Durante a fase de teste, é simulada uma cadeia de Markov utilizando as probabilidades de cada transição, como no Leap Graph, e considerando a primeira posição de cada usuário como sendo sua posição inicial em $\mathcal{D}^{teste}$.

3.3.2 MobDatU-Contact

O MobDatU-Contact estende o MobDatU para considerar não somente as popularidades das regiões e as transições entre elas mas também os contatos entre os usuários. Um contato entre dois usuários pode ser definido como qualquer tipo de interação realizada entre eles, que serve como evidência que tais usuários se conhecem (seja no mundo real ou virtual) e que, portanto, os deslocamentos de um podem influenciar na mobilidade de outro [Wang & Song, 2015]. Tais interações podem ser inferidas a partir de chamadas de telefonia móvel que são realizadas entre dois usuários, ou seja, se um usuário ligou para outro. No caso de dados de aplicativos georreferenciados, como o Twitter, um contato entre dois usuários pode ser definido a partir dos laços de seguidor e seguido, ou seja, dois usuários poderão ser considerados contatos se um usuário segue e/ou é seguido pelo outro. Além disso, é possível utilizar a intensidade dessas interações. Por exemplo, dois usuários poderiam ser considerados contatos somente se eles participarem de pelo menos x de chamadas. As estratégias usadas para definir um contato entre dois usuários serão apresentadas na Seção 4.2.

O MobDatU-Contact utiliza o mesmo processo do MobDatU para definir as regiões. Ele também cria um grafo de transições entre regiões. Porém, o cálculo das probabilidades de transições leva em consideração também a probabilidade de encontrar um usuário em uma dada região r_i no instante t, dado que pelo menos n de seus contatos estão presentes na mesma região no mesmo instante t ($P(L_{i,j,t}|C_{i,j,t}^{\geq n})$). Esta probabilidade é calculada nos dados de treino $\mathcal{D}^{treino}$. O MobDatU-Contact é motivado por resultados em [Nguyen & Szymanski, 2012] que mostram que ao considerar os contatos entre usuários conseguiram melhorar a previsão de mobilidade humana.

A probabilidade de encontrar o usuário em uma região é calculada como segue. Sejam os seguintes eventos:

- $L_{i,j,t}$: usuário u_i está localizado na região r_j na janela de tempo t
- $C^{\geq n}_{i,j,t}$: pelo menos n contatos do usuário i estão na região r_j na janela de tempo t
- $C^{k}_{i,j,t}$: exatamente k contatos do usuário i estão localizados na região r_j na janela de tempo t

Com isso nós definimos que a probabilidade $P(L_{i,j,t}|C^{\geq n}_{i,j,t})$ é dada pela probabilidade do evento $L_{i,j,t}$ acontecer ao mesmo tempo do evento $C^{\geq n}_{i,j,t}$ dividido pela probabilidade de pelo menos n dos contatos de u_i estarem na região r_j no tempo t. Esse cálculo é mostrado na equação seguinte.

$$P(L_{i,j,t}|C^{\geq n}_{i,j,t}) = \frac{P(L_{i,j,t} \ \wedge \ C^{\geq n}_{i,j,t})}{P(C^{\geq n}_{i,j,t})};$$

O denominador da equação anterior ($P(C^{\geq n}_{i,j,t})$) é calculado a partir do somatório das probabilidades de que exatamente k contatos de u_i estão em r_j no tempo t, para k variando de n ao numero de contatos de u_i, ou seja, k.

$$P(C^{\geq n}_{i,j,t}) = \sum_{k=n}^{|C_i|} P(C^{k}_{i,j,t});$$

O cálculo de $P(C^{k}_{i,j,t})$ é dado pela fração de janelas de tempo em que exatamente k contatos de u_i estavam presentes na região r_j. Se t_{max} é o número de janelas de tempo em $\mathcal{D}^{treino}$, $P(C^{k}_{i,j,t})$ é dado por:

$$P(C^{k}_{i,j,t}) = \frac{\sum_{t=1}^{t_{max}} \ (C^{k}_{i,j,t})}{t_{max}};$$

onde (condição)$= \begin{cases} 1, & \text{Se condição for verdadeira;} \\ 0, & \text{Caso contrário.} \end{cases}$

Já a probabilidade $\frac{P(L_{i,j,t} \ \wedge \ C^{\geq n}_{i,j,t})}{P(C^{\geq n}_{i,j,t})}$ é calculada a partir do somatório das probabilidades do evento $L_{i,j,t}$ acontecer ao mesmo tempo que o evento $C^{k}_{i,j,t}$ para todos os valores de k, ou seja, $\sum_{k=n}^{|C_i|} P(L_{i,j,t} \ \wedge \ C^{k}_{i,j,t})$.

$$P(L_{i,j,t} \ \wedge \ C^{\geq n}_{i,j,t}) = \sum_{k=n}^{|C_i|} P(L_{i,j,t} \ \wedge \ C^{k}_{i,j,t});$$

Para um dado valor de k, a probabilidade $P(L_{i,j,t} \wedge C^k_{i,j,t})$ é dada pela fração das janelas de tempo em que os eventos $L_{i,j,t}$ e C_{i,j,t_i} acontecem simultaneamente, ou seja:

$$P(L_{i,j,t} \wedge C^k_{i,j,t}) = \frac{\sum_{t=1}^{t_{max}}((L_{i,j,t}) \times (C^k_{i,j,t}))}{t_{max}};$$

Após calcular a probabilidade $P(L_{i,j,t}|C^{\geq n}_{i,j,t})$, ela será multiplicada pela popularidade da região e pelo peso atual das arestas para ser definida como novo peso das arestas. Ou seja, o mesmo cálculo que é mostrado na Figura 3.10 é realizado, porém com o acréscimo da multiplicação da probabilidade de encontrar um usuário em uma região dado que pelo menos n de seus contatos estava lá.

As probabilidades $P(L_{i,j,t}|C^{\geq n}_{i,j,t})$ são utilizadas no cálculo dos pesos das arestas do grafo de transições. Nós consideramos quatro diferentes abordagens para incorporar essas probabilidades no cálculo desses pesos. São elas:

1. *Transição e Contatos (TC)*: o peso de uma aresta (r_i,r_j) é computado pelo produto da frequência de transição entre regiões pela probabilidade $P(L_{i,j,t}|C^{\geq n}_{i,j,t})$, desconsiderando a popularidade da região r_j;

2. *Popularidade e Contatos (PC)*: o peso de uma aresta (r_i,r_j) é dada pelo produto da popularidade de r_j pela probabilidade $P(L_{i,j,t}|C^{\geq n}_{i,j,t})$, desconsiderando a frequência de transição;

3. *Apenas Contatos (C)*: o peso de uma aresta (r_i,r_j) é dado simplesmente pela probabilidade $P(L_{i,j,t}|C^{\geq n}_{i,j,t})$;

4. *Transição, Popularidade e Contatos (TPC)*: o peso de uma aresta (r_i,r_j) é dado pelo produto das três variáveis, a saber, frequência de transição entre duas regiões, popularidade de r_j e probabilidade $P(L_{i,j,t}|C^{\geq n}_{i,j,t})$.

Nos casos 1, 2 e 4 os pesos são posteriormente normalizados para representarem probabilidades de transição. Em avaliações preliminares, a estratégia 4 foi a que produziu melhores resultados, conforme será mostrado na Seção 5.2.

Durante a fase de teste, são simuladas duas cadeias de Markov. A primeira considera as três variáveis e é utilizada para os usuários para os quais foram identificados contatos em $\mathcal{D}^{teste}$. Já a segunda considera apenas a popularidade da região de destino e a frequência de transição, e é utilizada para os usuários para os quais não foram identificados contatos no conjunto de teste. Em ambas as cadeias de Markov, é considerada a primeira posição de cada usuário como sendo sua posição inicial em $\mathcal{D}^{teste}$.

3.4 Considerações Finais

Nesse capítulo, dois modelos estado-da-arte foram apresentados. O SMOOTH tem como premissa considerar que as pessoas tendem a ir para regiões populares e que regiões populares tendem a ficar mais populares. Já o Leap Graph parte da premissa que pessoas tendem a seguir a mesma trajetória durante o dia e que isso pode ser representado a partir de um grafo, no qual os vértices são as regiões que as pessoas visitam e as arestas as suas trajetórias.

Além dos modelos estado-da-arte, dois novos modelos de previsão de mobilidade humana foram propostos. O MobDatU que tem como premissa que pessoas tendem a seguir a mesma trajetória durante o dia, mas também que regiões populares influenciam em sua movimentação. O segundo modelo, o MobDatU-Contact, extende os princípios do MobDatU e acrescenta a premissa de que as relações de contato entre duas pessoas influencia na mobilidade. Ou seja, uma pessoa tende a estar na mesma região em que seus contatos também estejam.

Esses quatro modelos de previsão de mobilidade humana foram avaliados para fontes de dados distintas. As coleções de dados e a metodologia de avaliação são explicadas no Capítulo 4.

Capítulo 4

Metodologia de Avaliação

Nessa capítulo serão apresentadas as coleções de dados (Seção 4.1), bem como a metodologia adotada na avaliação dos modelos (Seção 4.2).

4.1 Coleções de Dados

Para avaliar os modelos de previsão foram utilizados dados de telefonia móvel e dados coletados do Twitter. Os dados de chamadas foram fornecidos por uma grande companhia de telefonia móvel brasileira e correspondem às chamadas realizadas em algumas cidades brasileiras durante intervalos de tempo pré-especificados. Os dados são anonimizados e contêm as seguintes informações para cada chamada:

- *Identificador da Chamada*: campo numérico que permite identificar unicamente cada chamada;
- *Identificador do Usuário*: campo numérico que permite identificar unicamente, mas de forma anonimizada, o usuário que realizou a chamada;
- *Hora Inicial*: hora de início da chamada;
- *Hora Final*: hora de término da chamada;
- *Antena inicial*: coordenadas geográficas (*i.e.* latitude e longitude) da antena à qual foi associada a chamada do usuário; corresponde à antena mais próxima do local onde o usuário se encontrava quando ele iniciou a chamada;
- *Antena final*: latitude e longitude da antena a qual foi associada a chamada do usuário; corresponde à antena mais próxima do local onde o usuário se encontrava quando ele terminou a chamada.

Já os dados do Twitter consistem em *tweets* georreferenciados, ou seja, *tweets* com coordenadas geográficas. Eles foram coletados pela *Stream API* da aplicação Twitter utilizando o filtro *location* que restringe a área de coleta para uma determinada região. A coleta pela API é realizada em tempo real. Logo, após a companhia de telefonia móvel informar os períodos de tempo dos dados que seriam fornecidos, a coleta de *tweets* foi planejada para os mesmos locais e períodos de tempo. Para cada *tweet* coletado, se tem:

- *Identificador do Tweet*: id do *tweet*;
- *Identificador do Usuário*: permite identificar, de forma anonimizada, o usuário que postou o *tweet*;
- *Latitude*: coordenada geográfica latitudinal do local onde o usuário se encontrava quando postou o *tweet*;
- *Longitude*: coordenada geográfica longitudinal do local onde o usuário se encontrava quando postou o *tweet*;
- *Hora*: hora em que foi postado o *tweet*;
- *Retweets*: lista de usuários (identificados pelos seus ids) que postaram *retweets* de mensagens previamente postadas pelo usuário.

Além disso para cada usuário se tem:

- *Seguidores*: lista de usuários (identificados pelos seus ids) que seguem o usuário;
- *Seguidos*: lista de usuários (identificados pelos seus ids) seguidos pelo usuário.

Os dois conjuntos de dados foram coletados de forma independente. Assim, não é possível identificar o mesmo usuário em coleções de fontes distintas. Portanto, os usuários das coleções de chamadas são considerados diferentes dos usuários das coleções do Twitter.

Os dados coletados foram filtrados para retirar todos os usuários que realizaram somente uma chamada ou postaram somente um *tweet*. Tais usuários foram considerados sem mobilidade e portanto descartados.

A Tabela 4.1 apresenta um sumário das coleções, após filtragem. Ela fornece para cada coleção, o local (cidade) e o período avaliado, os números de chamadas e de *tweets* assim como os respectivos números de usuários. No total, as coleções cobrem 5

Tabela 4.1: Coleções de Dados Utilizadas

Coleções	Data	Intervalo de Tempo	Chamadas		*Tweets*	
			# Chamadas	# Usuários	#*Tweets*	#Usuários
Belo Horizonte	21/10/2011	13h-21h	31.705	12.237	32.334	11.231
Belo Horizonte	31/12/2011	20h-04h	201.212	100.021	210.001	105.000
Belo Horizonte	03/01/2012	20h-04h	12.145	5.246	40.234	17.342
Belo Horizonte	03/02/2013	13h-20h	69.227	30.033	30.765	10.338
Belo Horizonte	10/03/2013	13h-20h	15.794	7.585	27.340	12.845
Belo Horizonte	02/03/2013	12h-19h	15.630	9.354	14.332	4.870
Belo Horizonte	22/06/2013	13h-21h	4.050	1.998	30.103	12.540
Belo Horizonte	26/06/2013	13h-21h	6.264	2.987	29.934	11.532
Belo Horizonte	11/09/2013	17h-23h	14.023	4.532	15.635	5.103
Fortaleza	29/06/2014	14h-21h	7.185	2.372	13.453	4.236
Recife	31/12/2011	20h-04h	21.123	10.000	45.321	20.192
Recife	03/01/2012	20h-04h	8.769	4.390	7.839	2.987
Recife	29/06/2014	14h-21h	13.335	4.923	13.577	3.981
Rio de Janeiro	28/08/2011	14h-20h	67.627	28.027	38.091	13.227
Rio de Janeiro	30/10/2011	14h-20h	58.610	25.593	37.931	12.498
Rio de Janeiro	04/12/2011	14h-20h	77.869	30.597	39.239	12.945
Rio de Janeiro	11/12/2011	14h-20h	56.159	23.563	40.123	13.002
Rio de Janeiro	31/12/2011	20h-04h	36.354	13.918	21.021	3.211
Rio de Janeiro	03/01/2012	20h-04h	20.231	9.134	45.322	19.443
Rio de Janeiro	29/03/2012	18h-22h	31.166	12.305	45.030	15.302
Rio de Janeiro	08/07/2012	14h-20h	7.579	3.384	30.213	13.490
Rio de Janeiro	27/11/2013	18h-00h	17.009	6.192	32.940	13.834
Rio de Janeiro	29/06/2014	14h-21h	5.120	1.132	14.033	3.643
Rio de Janeiro	13/07/2014	14h-21h	5.340	1.038	15.860	4.572
São Paulo	04/02/2012	15h-22h	3.370	1.159	25.370	11.930
São Paulo	25/11/2012	12h-18h	22.752	11.235	28.042	13.220
São Paulo	24/03/2013	13h-20h	44.499	20.787	50.323	20.334
Total			874.147	383.742	974.406	392.848

cidades, 21 dias diferentes, 874.147 chamadas realizadas por 383.742 usuários e 974.406 *tweets* postados por 392.848 usuários.

Note que o volume de *tweets* é superior ao número de chamadas em quase todas as coleções, com exceção de dois dias para Belo Horizonte e cinco dias para o Rio de Janeiro. Note também a variação na cobertura de usuários. Por exemplo, para Belo Horizonte (02/03/13), o número de usuários é duas vezes maior na coleção de chamadas. Já para São Paulo, o número de usuários é maior na coleção de *tweets* em dois dias analisados. Estas diferenças podem ser explicadas por aspectos socioculturais e eventos específicos que ocorreram em cada cidade nos períodos monitorados. Tais diferenças ilustram que pode ser vantajoso desenvolver modelos de previsão de mobilidade que se adequam bem a diferentes fontes de dados, uma vez que a cobertura de usuários por cada fonte pode variar entre locais (ou períodos) diferentes. Um exemplo disso é o posicionamento das antenas para os dados de chamadas. Ou seja, regiões que tendem a ter mais chamadas realizadas também possuem mais antenas para suportar a quantidade de chamadas. Isso, pode ser visto na Figura 4.1[1], que mostra a posição aproximada de cada antena para a cidade de Belo Horizonte e que é possível visualizar

[1]Mapa retirado do site da Telebrasil: http://www.telebrasil.org.br/panorama-do-setor/mapa-de-erbs-antenas

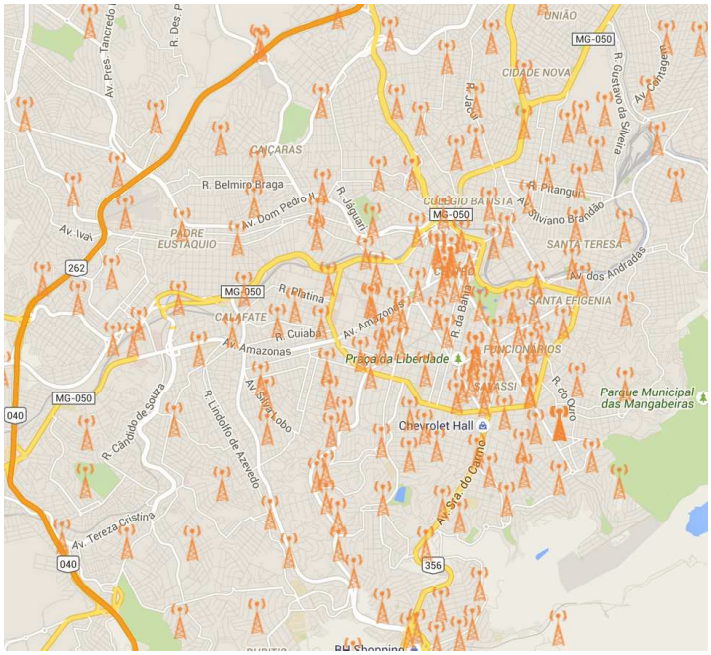


Figura 4.1: Localização das Antenas em Belo Horizonte. Fonte: Telebrasil

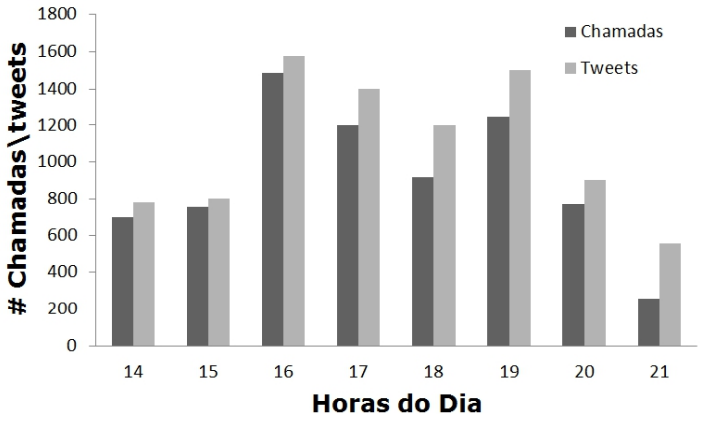


Figura 4.2: Quantidade de Chamadas/*Tweets* por Hora - Rio de Janeiro 29/06/14

uma maior densidade de antenas em algumas regiões do que outras.

As Figuras 4.2 e 4.3 mostram os números de chamadas, *tweets* e usuários ao longo do tempo na coleção do Rio de Janeiro (29/06/2014). Estas figuras ilustram que o volume de dados disponível varia ao longo do tempo. Por exemplo, para o período mostrado nas figuras, há picos nos números de chamadas/*tweets* e de usuários às 16h e às 19h. Tais variações são esperadas e podem refletir diretamente nos padrões de movimentação dos usuários. Por exemplo, os padrões de movimento às 16h podem ser diferentes daqueles observados às 21h. Mais ainda, a varação temporal no volume de dados disponível, que foi observada em todas as coleções, também pode afetar a precisão dos modelos aprendidos a partir de tais dados. Essa observação motiva

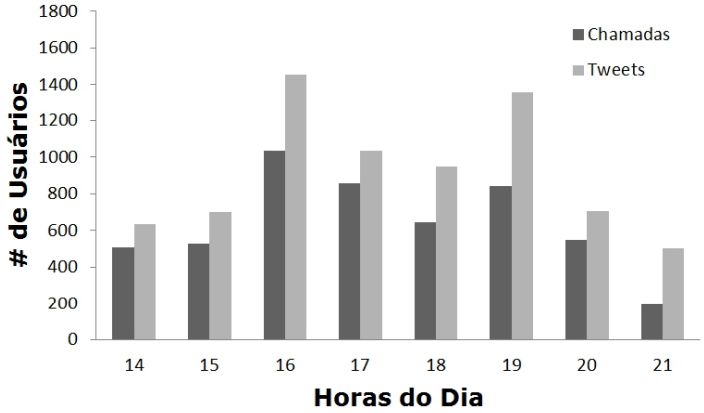


Figura 4.3: Quantidade de Usuários por Hora - Rio de Janeiro 29/06/14

a metodologia para aprender e avaliar os modelos de previsão. Esta metodologia é discutida na próxima seção.

4.2 Cenários de Avaliação

A métrica de avaliação adotada é a taxa de acerto que corresponde à fração das tuplas $< u_i, r_i, t >$ no conjunto de teste $\mathcal{D}^{teste}$ para as quais a previsão feita está correta.

A avaliação em si foi feita da seguinte forma. Cada coleção foi subdividida em intervalos de uma hora. Dadas as variações temporais nos volumes de dados observadas (Figura 4.2 e 4.3), optou-se por desenvolver um modelo de previsão para cada hora de forma a melhor capturar os padrões de mobilidade em diferentes períodos de tempo. Em seguida, os dados de cada subcoleção foram divididos entre treino $\mathcal{D}^{treino}$ e teste $\mathcal{D}^{teste}$. Duas estratégias foram adotadas para fazer esta divisão.

Para as coleções de dados que cobrem o mesmo período de tempo e mesmo dia da semana, mas semanas diferentes, o treinamento foi feito usando os dados de um dia ($\mathcal{D}^{treino}$), e a avaliação foi realizada utilizando os dados de um dia posterior ($\mathcal{D}^{teste}$). Esta estratégia, mostrada na Figura 4.4, foi adotada para as seguintes coleções:

- Belo Horizonte 31/12/2011 e Belo Horizonte 03/01/2012;
- Belo Horizonte 03/02/2013 e Belo Horizonte 10/03/2013;
- Recife 31/12/2011 e Recife 03/01/2012;
- Rio de Janeiro 28/08/2011 e Rio de Janeiro 30/10/2011;
- Rio de Janeiro 04/12/2011 e Rio de Janeiro 11/12/2011;

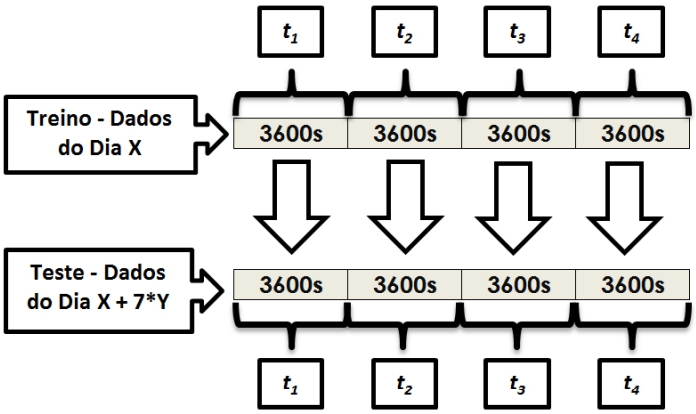


Figura 4.4: Separação dos dados em Treino e Teste - Dois Dias

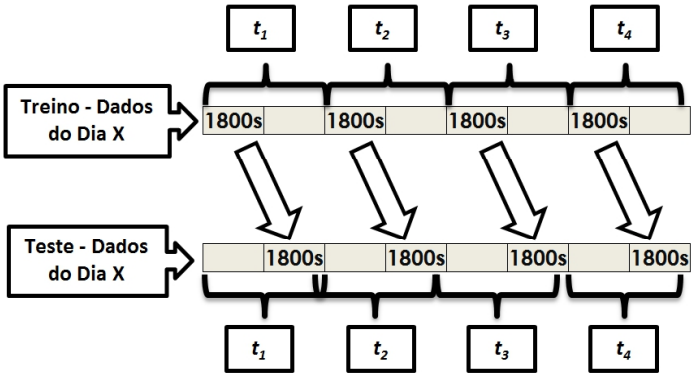


Figura 4.5: Separação dos dados em Treino e Teste - Um dia

- Rio de Janeiro 31/12/2011 e Rio de Janeiro 03/01/2012;
- Rio de Janeiro 29/06/2014 e Rio de Janeiro 13/07/2014.

Para as demais coleções, como não tivemos acesso a dados de múltiplos dias (mesmo dia de semana em semanas diferentes) cobrindo o mesmo período, foi adotada uma estratégia diferente. Cada hora no intervalo de tempo coberto pela coleção foi dividida em dois períodos de meia hora. O primeiro período foi utilizado para aprender o modelo (treino), e o segundo para avaliá-lo (teste). Esta estratégia é ilustrada na Figura 4.5. Para ambas as estratégias de divisão dos dados em treino e teste, um modelo foi treinado e avaliado para cada hora do período coberto pela coleção, conforme já mencionado.

Para cada coleção, cada modelo de mobilidade foi avaliado usando, como conjuntos de treino e teste, somente as chamadas, somente os *tweets* e tanto chamadas quanto *tweets*. O objetivo do último cenário é avaliar o desempenho dos modelos quando configurados com dados heterogêneos de forma conjunta. A melhor forma de combinar os dados das duas fontes não é óbvia, já que modelos diferentes podem ter desempe-

nhos relativos diferentes, dependendo do dado de entrada. Assim, nós consideramos e avaliamos duas estratégias de combinação de chamadas e *tweets*:

- Associação de *tweets* a chamadas: cada *tweet* é associado à antena mais próxima de sua localização;

- Associação de chamadas a *tweets*: cada antena da coleção de chamadas é considerada um ponto na região de simulação.

Um aspecto importante da avaliação é a definição das regiões R de cada modelo. Para chamadas, o Leap Graph e o SMOOTH consideram a localização de cada antena que aparece no conjunto $\mathcal{D}^{treino}$ como uma região r_i. Porém, como discutido na Seção 3.2.1, no caso do SMOOTH, novas regiões podem surgir durante o treino, de acordo com a movimentação dos usuários. Para os *tweets*, o SMOOTH considera inicialmente cada localização distinta associada a um *tweet* em $\mathcal{D}^{treino}$ como uma região, e novas regiões podem surgir durante o treino. Para o Leap Graph, os *tweets* foram agrupados em regiões circulares (com raio d) considerando suas localizações. Para os dados combinados, foram adotadas as mesmas abordagens dependendo da estratégia de combinação feita. Para o caso de um usuário estar dentro do raio de duas ou mais regiões, é considerado que ele está associado a região, cujo centro está mais próximo de sua localização. Já o MobDatU e o MobDatU-Contact sempre dividem a área total de cada cidade em regiões quadrangulares não sobrepostas (vide Seção 3.3.1), independentemente do tipo de dado.

Em todos os casos, consideramos a distância d que define cada região como 500 metros, que é o raio de cobertura de uma antena. Este valor foi escolhido para que os resultados dos vários cenários sejam comparáveis[2], uma vez que para o Leap Graph, quando aprendido a partir de dados de telefonia móvel, cada região é determinada pelo raio de cobertura de uma antena.

Em relação ao MobDatU-Contact, é necessário determinar como os contatos de um usuário serão inferidos dos dados. As coleções de telefonia móvel disponíveis contêm somente o identificador do usuário que realizou a chamada. O receptor da chamada não é identificado. Logo, não é possível extrair evidências de contatos entre os usuários a partir desses dados. Sendo assim, optou-se por avaliar o MobDatU-Contact usando apenas os dados do Twitter.

[2]Note que as áreas das regiões do SMOOTH e do Leap Graph (πd^2) são maiores que as áreas das regiões definidas pelo MobDatU (d^2). Logo a nossa avaliação favorece os modelos de referência, uma vez que as previsões são mais difíceis para regiões com áreas menores.

Como discutido na Seção 3.3.2, existem várias maneiras de inferir o que seriam os contatos de um usuário a partir dos *tweets*. Optou-se por explorar duas estratégias[3]. A primeira, que já foi abordada na literatura [Musolesi & Mascolo, 2007; Noulas et al., 2012b; Noulas & Mascolo, 2013; Scellato et al., 2011], consiste em utilizar os laços de seguidores e seguidos. Ou seja, se um usuário u_i segue e é seguido por um usuário u_j, ambos são considerados contatos um do outro. Esta estratégia foi nomeada por *Seguidor-Seguido*.

A segunda estratégia, nomeada de *Retweets*, consiste em considerar a intensidade dos retweets realizados pelos usuários. Seja rt_{ij} o número de vezes que o usuário u_i postou um *retweet* de alguma mensagem de u_j. A intensidade dos retweets trocados entre u_i e u_j é dada por

$$I_{ij} = \frac{rt_{ij} + rt_{ji}}{\sum_{k=1,k\neq i}^{|U|} rt_{i,k} + \sum_{k=1,k\neq i}^{|U|} rt_{k,i}},$$

Ou seja, ela captura a fração de todos os *retweets* que envolvem o usuário u_i dos quais o usuário u_j participou.

Um usuário u_j é considerado um contato de u_i (e vice-versa) se $I_{i,j}$ for igual ou superior a um limiar θ. O Capítulo 5 apresenta avaliação do impacto da escolha de θ nas previsões.

[3]Note que em ambas as estratégias, a relação de contato é bidirecional; ou seja, se o usuário u_i é um contato de u_j, u_j também é contato de u_i

Capítulo 5

Resultados

Este capítulo apresenta os principais resultados da avaliação dos dois modelos propostos - MobDatU e MobDatU-Contact - bem como dos modelos de referência nos vários cenários considerados que foram discutidos no Capítulo 4. A Seção 5.1 foca no MobDatU, comparando-o com os modelos de referência. Já a Seção 5.2 estende a avaliação para incluir o novo modelo MobDatU-Contact.

5.1 MobDatU e Modelos de Referência

Nesta seção serão discutidos os resultados da avaliação do MobDatU e dos dois modelos de referência, o SMOOTH e o Leap Graph. Como discutido no Capítulo 4, nesta avaliação nós utilizamos um total de 27 coleções que cobrem 5 cidades diferentes. Essas coleções foram divididas em dois grupos. O primeiro grupo, com 14 coleções, em que o treino e o teste foram realizados em dias diferentes, conforme mostrado na linha do tempo da Figura 4.4. Já o segundo grupo com 13 coleções em que o treino e o teste foram realizados no mesmo dia, conforme mostrado na linha do tempo da Figura 4.5.

Os modelos foram avaliados para os quatro cenários que foram discutidos na Seção 4.2. Esses cenários consistem em utilizar os dados de telefonia móvel e do Twitter de forma conjunta e separada.

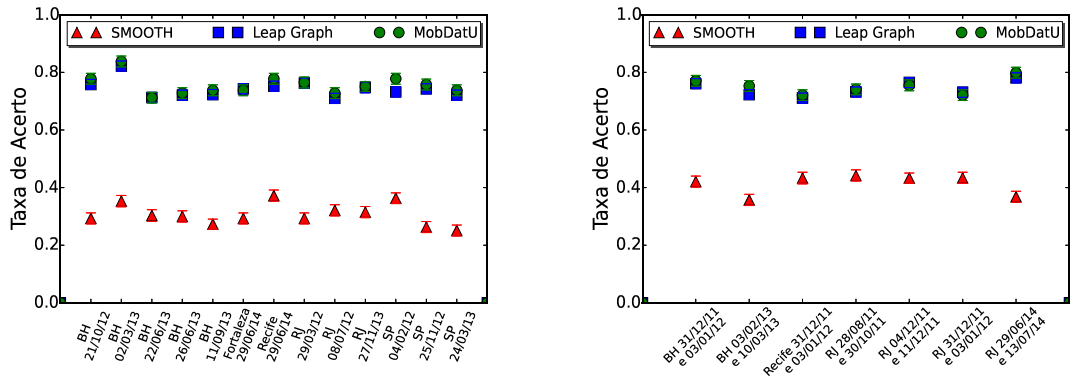
A Tabela 5.1 apresenta alguns resultados para os modelos de referência e para o MobDatU. Nessa tabela são mostrados para cada cenário, as taxas de acerto médias e intervalos de confianças de 95% em todas as horas do período de cada coleção. A Tabela 5.1 apresenta os resultados para duas coleções em que o treino e o teste foram realizados no mesmo dia, além de duas coleções em que o treino e o teste foram realizados em dias diferentes. Os melhores valores para cada cenário, incluindo empates estatísticos (com 95% de confiança) são mostrados em negrito.

Tabela 5.1: Avaliação do MobDatU e Modelos de Referência: taxas de acerto médias e intervalos de confiança de 95% (melhores resultados em negrito)

Rio de Janeiro Treino no dia 29/06/2014 e teste no dia 13/07/2014				
Modelo	Chamadas	*Tweets*	*Tweets* para Chamadas	Chamadas para *Tweets*
SMOOTH	0,368±0.0177	0.521±0.0189	0.481±0.0193	0.516±0.0189
Leap Graph	**0.781±0.0168**	0.448±0.0195	**0.738±0.0182**	0.416±0.0188
MobDatU	**0.799±0.0192**	**0.551±0.0201**	**0.744±0.0186**	**0.531±0.0164**
Belo Horizonte Treino no dia 03/02/2013 e teste no dia 10/03/2013				
Modelo	Chamadas	*Tweets*	*Tweets* para Chamadas	Chamadas para *Tweets*
SMOOTH	0.357±0.0190	0.519±0.0162	0.324±0.0221	0.532±0.0198
Leap Graph	0.723±0.0165	0.395±0.0166	0.748±0.0153	0.379±0.0215
MobDatU	**0.753±0.0168**	**0.565±0.0182**	**0.768±0.0187**	**0.571±0.0194**
Rio de Janeiro Treino e teste no dia 27/11/2013				
Modelo	Chamadas	*Tweets*	*Tweets* para Chamadas	Chamadas para *Tweets*
SMOOTH	0.273±0.0195	0.531±0.0181	0.391±0.0196	**0.534±0.0198**
Leap Graph	**0.748±0.0189**	0.441±0.0177	**0.721±0.0173**	0.351±0.0205
MobDatU	**0.751±0.0168**	**0.581±0.0183**	**0.730±0.0185**	**0.546±0.0198**
Belo Horizonte Treino e teste no dia 11/09/2013				
Modelo	Chamadas	*Tweets*	*Tweets* para Chamadas	Chamadas para *Tweets*
SMOOTH	0.273±0.0194	0.533±0.0162	0.258±0.0219	**0.576±0.0188**
Leap Graph	**0.723±0.0175**	0.365±0.0167	**0.761±0.0184**	0.418±0.0195
MobDatU	**0.738±0.0188**	**0.551±0.0181**	**0.751±0.0185**	**0.579±0.0178**

No geral, nota-se que cada modelo de referência funciona melhor se configurado com dados homogêneos para os quais ele foi avaliado anteriormente: o Leap Graph consegue taxa de acertos melhores quando utiliza somente chamadas e o SMOOTH consegue resultados melhores quando utiliza somente *tweets*. Especificamente, considerando a coleção de dados do Rio de Janeiro para treino e teste em dias diferentes (primeiro grupo de resultados da Tabela 5.1), nota-se que, quando os dados de chamadas são usados como entrada, o SMOOTH tem um desempenho médio degradado de 53% em relação ao Leap Graph. Já quando são usados os *tweets*, o Leap Graph possui um desempenho 14% pior que o SMOOTH, em média. Quanto aos cenários com dados heterogêneos, a taxa de acerto de cada modelo foi um pouco menor em relação ao seu melhor cenário. Ou seja, os modelos de referência se comportaram um pouco pior em cenários com dados heterogêneos.

Em contrapartida, o MobDatU tem um desempenho comparável aos melhores resultados dos modelos de referência em todos os cenários. Ainda na coleção do Rio de Janeiro, com treino no dia 29/06/2014 e teste no dia 13/07/2014, o MobDatU tem uma



(a) Treino e Teste no Mesmo Dia

(b) Treino e Teste em Dias Diferentes

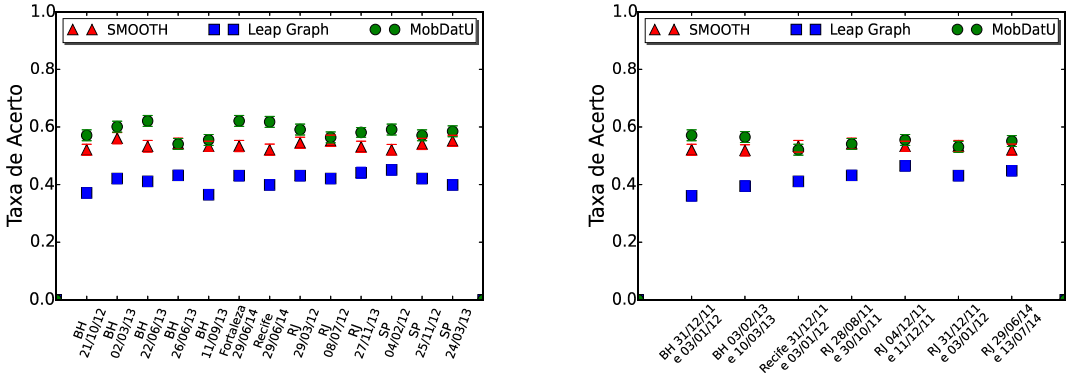
Figura 5.1: Taxas de Acerto Médias para o Cenários de Chamada para o MobDatU e Modelos de Referência - Todas Coleções

taxa de acerto média ligeiramente maior que a do Leap Graph e muito superior (109%) à do SMOOTH no cenário de chamadas. Já para o cenário de *tweets*, o MobDatU supera o Leap Graph em 18% e produz resultados similares aos do SMOOTH. Comparando os dois cenários com dados homogêneos, nota-se que o MobDatU tem uma taxa de acerto média bem maior no cenário de chamadas. A razão para este resultado é um maior número de regiões distintas presentes somente no conjunto de teste para os dados de *tweets*. Regiões que não aparecem no conjunto de treino terão popularidade e taxas de transição nulas no modelo e nunca serão previstas. Logo, regiões que só aparecem no teste necessariamente levam a previsões erradas do modelo[1].

Em relação ao cenário com dados heterogêneos, assim como os modelos de referência, o MobDatU apresenta um desempenho um pouco pior em relação aos cenários de dados homogêneos, mas ainda superior, em média, ao dos modelos de referência. Quanto às estratégias de combinação de dados, observa-se que cada modelo de referência teve melhor desempenho quando a estratégia adotada faz a associação tendo como alvo o tipo de dado para o qual o modelo tem melhor precisão (chamadas para o Leap Graph e para o MobDatU, *tweets* para o SMOOTH).

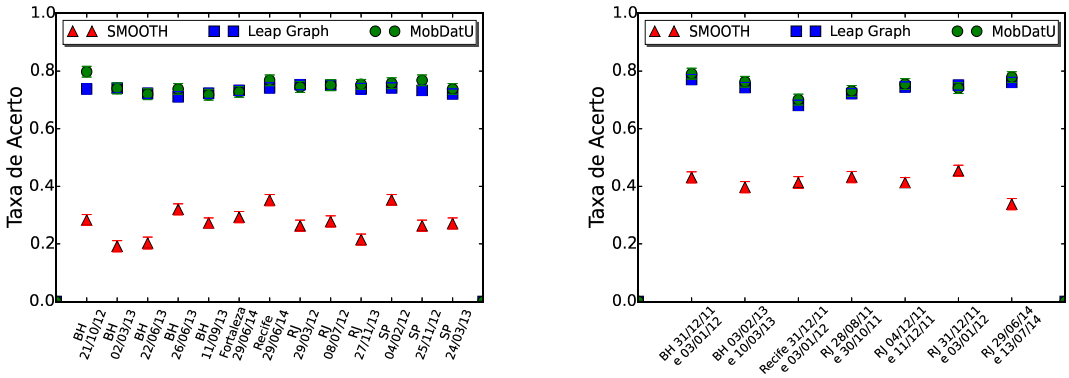
Os resultados obtidos para as outras três coleções de dados mostradas na Tabela 5.1 apresentam um comportamento similar aos discutidos acima. Entretanto, algumas exceções ocorreram nos cenários de dados heterogêneos. Por exemplo, nas duas coleções de Belo Horizonte, o MobDatU teve resultados piores nos cenários com dados homogêneos. Vale ressaltar que o uso de dados heterogêneos viabiliza uma realização de previsões para um número de usuários (isto é, cobertura de usuários) potencialmente

[1] Note que este maior número de regiões novas no teste foi observada em todas coleções, a despeito das diferenças de volumes de dados de chamadas e *tweets* discutidos na Seção 4.1.



(a) Treino e Teste no Mesmo Dia (b) Treino e Teste em Dias Diferentes

Figura 5.2: Taxas de Acerto Médias para o Cenários de *Tweets* para o MobDatU e Modelos de Referência - Todas Coleções



(a) Treino e Teste no Mesmo Dia (b) Treino e Teste em Dias Diferentes

Figura 5.3: Taxas de Acerto Médias para o Cenários de *Tweets* para Chamadas para o MobDatU e Modelos de Referência - Todas Coleções

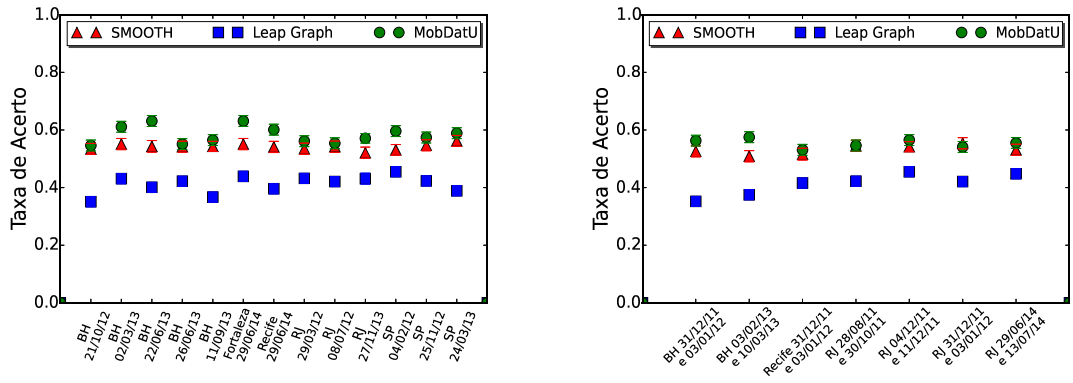
muito maior[2], o que pode compensar eventuais perdas nas taxas de acerto quando comparadas às obtidas com dados homogêneos.

As Figuras 5.1 - 5.4[3] mostram as taxas de acerto médias para os três modelos e para todas as coleções[4]. As mesmas conclusões discutidas acima são válidas para todos os cenários.

[2]Já que as coleções de chamadas e *tweets* não apresentam nenhum parâmetro que indique a existência de um mesmo usuário em ambas as fontes, foi considerado que os usuários dos dados de chamadas eram diferentes dos de *tweets* conforme descrito na Seção 4.1. Entretando, vale ressaltar que caso as coletas de dados sejam coordenadas, de forma a ser possível a identificação do mesmo usuário nas duas fontes de dados, tal simplificação não precisaria ser feita e poderia resultar em previsões ainda melhores e mais realistas.

[3]Todas as figuras apresentam o intervalo de confiança de 95%. Porém os intervalos são muito estreitos, o que dificulta a sua visualização.

[4]Note que como temos 7 cenários em que o treino e teste foram realizados com dados de dias diferentes, as figuras mostram resultados para 20 cenários distintos.



(a) Treino e Teste no Mesmo Dia

(b) Treino e Teste em Dias Diferentes

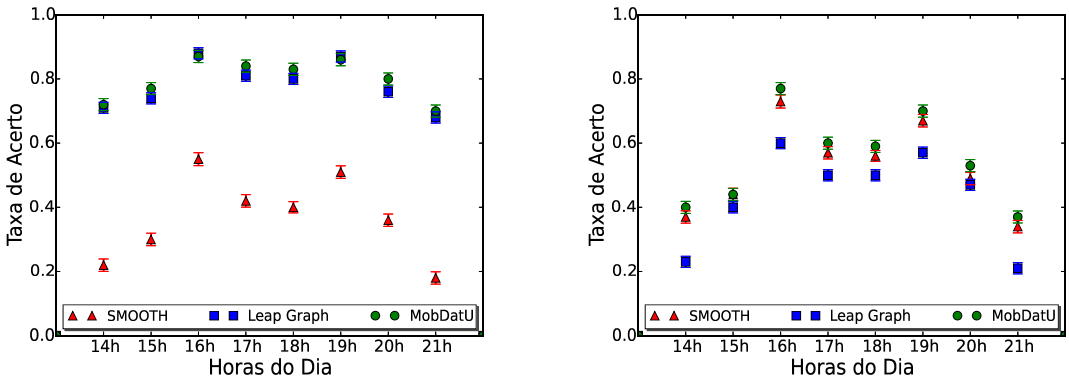
Figura 5.4: Taxas de Acerto Médias para o Cenários de Chamadas para *Tweets* para o MobDatU e Modelos de Referência - Todas Coleções

No geral, observamos que os ganhos do MobDatU sobre o SMOOTH e o Leap Graph chegaram a 67% e 31% em média, respectivamente. Mais ainda, quando comparado ao melhor modelo de referência em cada cenário, o MobDatU chegou a atingir ganhos médios de 9%.

As Figuras 5.5 e 5.6 mostram as taxas de acerto médias de cada modelo para cada intervalo de uma hora ao longo do período coberto pela coleção do Rio de Janeiro para os cenários com dados homogêneos (Figura 5.5) e com dados heterogêneos (Figura 5.6). Comparando estas figuras com a Figura 4.2, nota-se que as maiores taxas de acertos foram obtidas em períodos com maior volume de dados, o que era esperado. Um exemplo é o horário de 21h: conforme a Figura 4.2, o volume de dados disponível atinge um valor mínimo nesse período, o mesmo ocorre com as taxas de acerto dos três modelos. Conforme discutido na Seção 4.1 o volume de dados afeta a eficácia dos modelos. Ou seja, quanto maior o volume de dados existente para aprender os padrões de mobilidade humana mais preciso o modelo será.

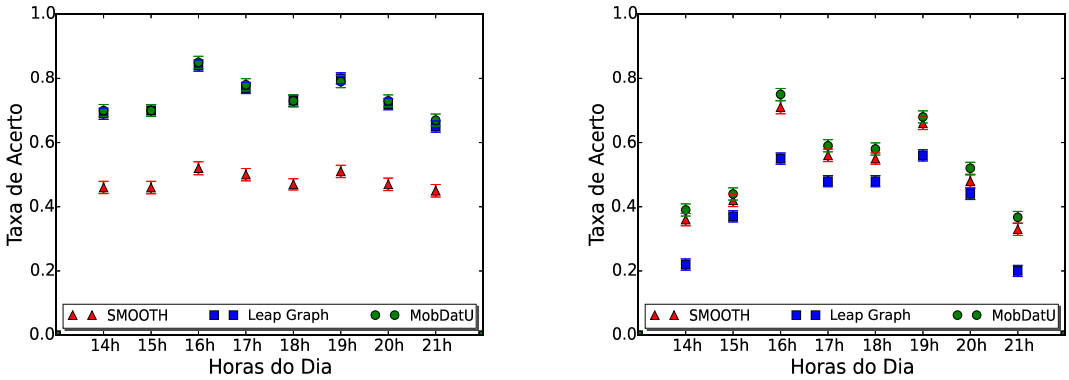
As Figuras 5.5 e 5.6 também mostram que os ganhos do MobDatU sobre o melhor modelo de referência podem ser maiores que os mostrados na Tabela 5.1. Por exemplo, na Figura 5.5(a), o MobDatU supera o Leap Graph em 5% , enquanto na Figura 5.5(b), para o intervalo de 16 horas, a taxa de acerto do MobDatU supera em 6% a do SMOOTH (melhor modelo de referência) para o período de 20 horas. Resultados semelhantes foram observados para todas as 27 coleções, conforme mostram os gráficos das Figuras A.1-A.38 no Apêndice A.

Em suma, os resultados mostram que o MobDatU foi capaz de se adequar bem a dados distintos, conseguindo desempenho comparável (e por vezes superior) ao do melhor modelo de referência, em todos os cenários com dados homogêneos e heterogê-



(a) Chamadas (b) *Tweets*

Figura 5.5: Taxas de Acertos Médias por Hora para os Cenários de Dados Homogêneos para o MobDatU e Modelos de Referência - Rio de Janeiro, treino no dia 29/06/2014 e teste no dia 13/07/2014



(a) *Tweets* para Chamadas (b) Chamadas para *Tweets*

Figura 5.6: Taxas de Acertos Médias por Hora para os Cenários de Dados Heterogêneos para o MobDatU e Modelos de Referência - Rio de Janeiro, treino no dia 29/06/2014 e teste no dia 13/07/2014

neos. Em contrapartida, tanto SMOOTH quanto Leap Graph podem ter grande perda de desempenho, dependendo do tipo de dado de entrada. Ou seja, o SMOOTH teve uma grande perda de desempenho para os cenários de chamadas (como mostrado nas Figuras 5.1 e 5.3) e o Leap Graph para os cenários de *tweets* (Figuras 5.2 e 5.4). A superioridade do nosso modelo pode ser explicada por ele considerar tanto a popularidade das regiões quanto a frequência de transições entre as regiões. Ambos os aspectos são fatores importantes que influenciam como as pessoas se movimentam em uma cidade. Além disto, como estes aspectos podem ser capturados tanto pelas chamadas telefônicas realizadas quanto pelos *tweets* compartilhados, o MobDatU mostrou um bom funcionamento para ambos os tipos de dados.

5.2 MobDatU-Contact e MobDatU

Esta seção apresenta os resultados do modelo MobDatU-Contact, comparando-o com o MobDatU. Como discutido na Seção 4.2, nós avaliamos o MobDatU-Contact apenas para os dados do Twitter. O MobDatU-Contact foi avaliado para todas as 27 coleções e, como na Seção 5.1, utilizamos a taxa de acerto como métrica de avaliação.

Antes de comparar o resultados obtidos pelos dois modelos propostos, nós analisamos o impacto de dois parâmetros do MobDatU-Contact na sua previsão, a saber a abordagem adotada para o cálculo da probabilidade de transição entre duas regiões e do número mínimo n de contatos do usuário a serem considerados no cálculo da probabilidade $P(L_{i,j,t}|C^{\geq n}_{i,j,t})$. Conforme apresentado na Seção 3.3.2 foram propostas quatro abordagens para o cálculo da probabilidade de transição. São elas: *Transição e Contatos (TC)*, *Popularidade e Contatos(PC)*, *Apenas Contatos (C)* e *Transição, Popularidade e Contatos (TPC)*.

Nós avaliamos a eficácia das quatro abordagens considerando as duas estratégias de identificação dos contatos de um usuário. A primeira estratégia chamada *Seguidor-Seguido* e a segunda chamada *Retweets*. Para essa última estratégia consideramos os valores do limiar θ iguais a 10%, 25%, 50% e 75%.

A Tabela 5.2 mostra as taxas de acerto médias e intervalos de confiança de 95% correspondentes ao MobDatU-Contact para as quatro abordagens utilizadas no cálculo da probabilidade de transição entre duas regiões. Nessa tabela nós avaliamos as abordagens para as mesmas quatro coleções apresentadas na Tabela 5.1 e consideramos parâmetro n, que define o número mínimo de contatos, igual a 1.

Utilizando, como exemplo, a coleção do Rio de Janeiro com treino no dia 29/06/2014 e teste no dia 13/07/2014, é possível perceber que a abordagem (C) produziu os piores resultados. Já os resultados para as abordagens (TC) e (PC) foram similares. Entretanto, estes não foram tão bons quanto os produzidos pela abordagem que considera as três variáveis (TPC), que supera em 4% o segundo melhor método. Com isso, é possível perceber que mesmo considerando os contatos de um usuário, a sua movimentação também sofre influência da popularidade das regiões e das frequências de transições entre regiões. Resultados semelhantes foram observados para três outras coleções presentes na Tabela 5.2 e também para as demais 23 coleções. Assim, nas avaliações subsequentes do MobDatU-Contact, nós consideramos a abordagem que utiliza as três variáveis.

Além das quatro abordagens para calcular a probabilidade de transição entre duas regiões, nós variamos o parâmetro n do MobDatU-Contact que define o número *mínimo* de contatos em uma região, utilizado no cálculo da probabilidade $P(L_{i,j,t}|C^{\geq n}_{i,j,t})$.

Tabela 5.2: Estratégias para o cálculo da probabilidade de transição entre regiões no modelo MobDaU-Contact considerando pelo menos um contato de um usuário (melhores resultados em negrito)

Rio de Janeiro Treino no dia 29/06/2014 e teste no dia 13/07/2014					
Estratégia		*Transição e Contatos (TC)*	*Popularidade e Contatos (PC)*	*Apenas Contatos (C)*	*Transição, Popularidade e Contatos (TPC)*
Seguidor-Seguido		0.507±0.0145	0.523±0.0177	0.342±0.0198	0.597±0.0178
Retweets	θ=10%	0.542±0.0152	0.572±0.0182	0.368±0.0175	0.638±0.0189
	θ=25%	**0.622±0.0169**	**0.653±0.0147**	**0.411±0.0149**	**0.688±0.0189**
	θ=50%	**0.635±0.0173**	**0.665±0.0162**	**0.421±0.0157**	**0.694±0.0177**
	θ=75%	0.612±0.0163	0.621±0.0149	0.394±0.0184	0.645±0.0187
Belo Horizonte Treino no dia 03/02/2013 e teste no dia 10/03/2013					
Estratégia		*Transição e Contatos (TC)*	*Popularidade e Contatos (PC)*	*Apenas Contatos (C)*	*Transição, Popularidade e Contatos (TPC)*
Seguidor-Seguido		0.504±0.0165	0.513±0.0174	0.339±0.0186	0.567±0.0208
Retweets	θ=10%	0.531±0.0142	0.567±0.0181	0.363±0.0167	0.598±0.0179
	θ=25%	**0.625±0.0178**	**0.641±0.0157**	**0.432±0.0145**	**0.678±0.0184**
	θ=50%	0.607±0.0153	0.625±0.0162	0.412±0.0188	0.644±0.0185
	θ=75%	0.601±0.0161	0.614±0.0159	0.383±0.0145	0.625±0.0184
Rio de Janeiro Treino e teste no dia 27/11/2013					
Estratégia		*Transição e Contatos (TC)*	*Popularidade e Contatos (PC)*	*Apenas Contatos (C)*	*Transição, Popularidade e Contatos (TPC)*
Seguidor-Seguido		0.513±0.0155	0.522±0.0184	0.338±0.0156	0.613±0.0169
Retweets	θ=10%	0.541±0.0133	0.559±0.0191	0.353±0.0187	0.645±0.0179
	θ=25%	**0.624±0.0169**	**0.631±0.0167**	**0.417±0.0155**	**0.675±0.0177**
	θ=50%	**0.633±0.0143**	**0.645±0.0152**	**0.429±0.0186**	**0.669±0.0193**
	θ=75%	0.613±0.0162	0.623±0.0151	0.392±0.0142	0.638±0.0185
Belo Horizonte Treino e teste no dia 11/09/2013					
Estratégia		*Transição e Contatos (TC)*	*Popularidade e Contatos (PC)*	*Apenas Contatos (C)*	*Transição, Popularidade e Contatos (TPC)*
Seguidor-Seguido		0.523±0.0165	0.522±0.0154	0.348±0.0158	0.573±0.0182
Retweets	θ=10%	0.531±0.0183	0.559±0.0181	0.363±0.0167	0.612±0.0191
	θ=25%	0.618±0.0179	0.628±0.0169	0.405±0.0175	0.655±0.0189
	θ=50%	**0.639±0.0153**	**0.645±0.0158**	**0.421±0.0185**	**0.673±0.0176**
	θ=75%	0.610±0.0166	0.611±0.0167	0.380±0.0146	0.625±0.0188

Considerando os critérios explorados pelas duas estratégias de identificação do contato (*Retweets* e *Seguidor-Seguido*), nós encontramos usuários com até três contatos nas coleções de treino.

Com isso, a Tabela 5.3 mostra os resultados para diferentes valores de n para quatro coleções, sendo duas com treino e teste em dias diferentes e duas com treino e teste em dias iguais. Para cada valor de n a tabela mostra resultados para as duas

Tabela 5.3: Estratégia de Definição de Contato (melhores resultados em negrito)

Rio de Janeiro Treino no Dia 29/06/2014 e Teste no Dia 13/07/2014				
Estratégia		$n = 1$	$n = 2$	$n = 3$
Seguidor-Seguido		0.597±0.0178	0.575±0.0137	0.536±0.0141
Retweets	θ=10%	0.638±0.0189	0.621±0.0132	0.613±0.0132
	θ=25%	**0.688±0.0189**	**0.657±0.0135**	**0.624±0.0142**
	θ=50%	**0.694±0.0177**	**0.645±0.0142**	
	θ=75%	0.645±0.0187		
Belo Horizonte Treino no Dia 03/02/2013 e Teste no Dia 10/03/2013				
Estratégia		$n = 1$	$n = 2$	$n = 3$
Seguidor-Seguido		0.567±0.0208	0.543±0.0145	0.524±0.0141
Retweets	θ=10%	0.598±0.0179	0.567±0.0142	0.543±0.0152
	θ=25%	**0.678±0.0184**	**0.651±0.0146**	**0.614±0.0143**
	θ=50%	0.644±0.0185	0.628±0.0152	
	θ=75%	0.625±0.0184		
Rio de Janeiro Treino e Teste no Dia 27/11/2013				
Estratégia		$n = 1$	$n = 2$	$n = 3$
Seguidor-Seguido		0.613±0.0169	0.583±0.0155	0.543±0.0142
Retweets	θ=10%	0.645±0.0179	0.627±0.0140	0.593±0.0152
	θ=25%	**0.675±0.0177**	**0.656±0.0144**	**0.623±0.0144**
	θ=50%	**0.669±0.0193**	**0.648±0.0154**	
	θ=75%	0.638±0.0185		
Belo Horizonte Treino e Teste no Dia 11/09/2013				
Estratégia		$n = 1$	$n = 2$	$n = 3$
Seguidor-Seguido		0.573±0.0182	0.552±0.0145	0.527±0.0147
Retweets	θ=10%	0.612±0.0191	0.596±0.0149	0.565±0.0172
	θ=25%	0.655±0.0189	0.633±0.0149	**0.602±0.0144**
	θ=50%	**0.673±0.0176**	**0.659±0.0164**	
	θ=75%	0.625±0.0188		

estratégias de identificação de contato. Note que para a estratégia de *Retweet* com $\theta = 50\%$, o número máximo de contatos possível de um usuário é dois, uma vez que o uso deste limiar exige que para ser um contato de um usuário u_i, um usuário u_j deve participar de pelo menos 50% dos *retweets* que envolvem u_i. Pela mesma razão, para $\theta = 75\%$, um usuário pode ter no máximo um contato.

A Tabela 5.3 mostra que os melhores resultados foram obtidos ao considerar o valor de $n = 1$ e as estratégia de *Retweets* com θ igual a 25% ou 50%. O mesmo foi observado para todas as coleções. Investigando o modelo, nós percebemos que quanto maior a probabilidade de encontrar um usuário em uma região dado que seus contatos também estão na mesma região, maior é a taxa de acerto do MobDatU-Contact. A razão dos melhores resultados da estratégia de *Retweets* com θ igual a 25% ou 50% será discutida mais adiante.

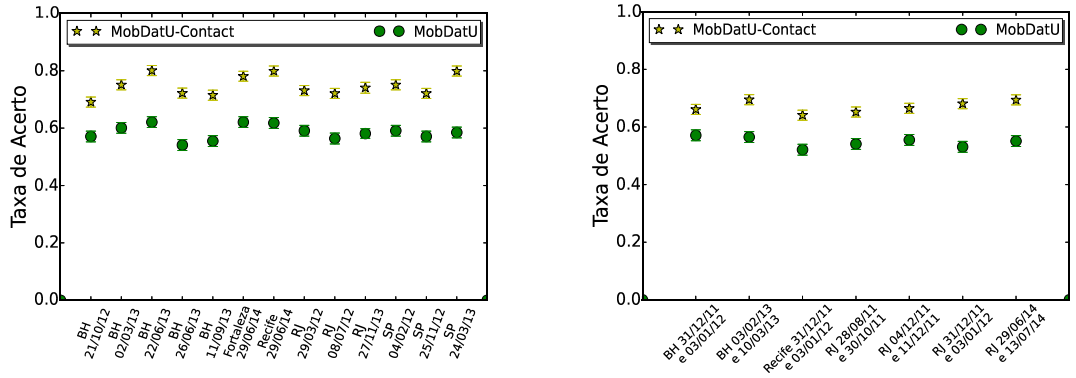
Tendo estudado o impacto dos principais parâmetros do MobDatU-Contact no seu desempenho, nós comparamos os seus resultados com os obtidos pelo MobDatU. A Tabela 5.4 mostra os resultados do MobDatU-Contact com a melhor estratégia de

Tabela 5.4: Avaliação do MobDatU-Contact e MobDatU: Taxas de Acerto Médias e Intervalos de Confiança de 95% (melhores resultados em negrito)

Rio de Janeiro **Treino no dia 29/06/2014 e teste no dia 13/07/2014**	
Modelo	*Tweets*
MobDatU	0.551±0.0201
MobDatU-Contact *Seguidor-Seguido*	0.597±0.0178
MobDatU-Contact θ=10%	0.638±0.0189
MobDatU-Contact θ=25%	**0.688±0.0189**
MobDatU-Contact θ=50%	**0.694±0.0177**
MobDatU-Contact θ=75%	0.645±0.0187
Belo Horizonte **Treino no dia 03/02/2013 e teste no dia 10/03/2013**	
Modelo	*Tweets*
MobDatU	0.545±0.0182
MobDatU-Contact *Seguidor-Seguido*	0.567±0.0208
MobDatU-Contact θ=10%	0.598±0.0179
MobDatU-Contact θ=25%	**0.678±0.0184**
MobDatU-Contact θ=50%	0.644±0.0185
MobDatU-Contact θ=75%	0.625±0.0184
Rio de Janeiro **Treino e teste no dia 27/11/2013**	
Modelo	*Tweets*
MobDatU	0.581±0.0183
MobDatU-Contact *Seguidor-Seguido*	0.613±0.0169
MobDatU-Contact θ=10%	0.645±0.0179
MobDatU-Contact θ=25%	**0.675±0.0177**
MobDatU-Contact θ=50%	**0.669±0.0193**
MobDatU-Contact θ=75%	0.638±0.0185
Belo Horizonte **Treino e teste no dia 11/09/2013**	
Modelo	*Tweets*
MobDatU	0.551±0.0181
MobDatU-Contact *Seguidor-Seguido*	0.573±0.0182
MobDatU-Contact θ=10%	0.612±0.0191
MobDatU-Contact θ=25%	0.655±0.0189
MobDatU-Contact θ=50%	**0.673±0.0176**
MobDatU-Contact θ=75%	0.625±0.0188

cálculo da probabilidade de transição e com $n = 1$, assim como os resultados do MobDatU para quatro de nossas coleções. O MobDatU-Contact supera significativamente os resultados obtidos pelo MobDatU. Mesmo as estratégias que tiveram o pior resultado (*Seguidor-Seguido* e *Retweets* com $\theta = 75\%$), estes foram melhores que o MobDatU. Por exemplo, para a coleção do Rio de Janeiro com treino e teste no dia 27/11/2013 (terceiro grupo de resultados da tabela), o MobDatU-Contact produziu ganhos em termos de taxa de acerto média sobre o MobDatU de 5% a 14%.

Como discutido anteriormente, os melhores resultados para o Rio de Janeiro com treino e teste no dia 27/11/2013 foram os valores de θ igual a 25% e 50%. Isso também foi observado nas demais coleções apresentadas na Tabela 5.4, no qual o melhor resultado foi considerando θ igual a 25% (por exemplo a coleção de Belo Horizonte com



(a) Treino e Teste no Mesmo Dia

(b) Treino e Teste em Dias Diferentes

Figura 5.7: Taxas de Acerto Médias para o MobDatU e MobDatU-Contact na Melhor Configuração - Todas Coleções

treino no dia 03/02/2013 e teste no dia 10/03/2013) ou considerando θ igual a 50% (por exemplo a coleção de Belo Horizonte com treino e teste no dia 11/09/2013). Em alguns casos não houve diferença significativa entre ambos os valores de θ (coleção do Rio de Janeiro para treino e teste em dias diferentes). Um sumário dos resultados para todas as coleções pode ser observado na Figura 5.7. Essa figura mostra que em todas as coleções a melhor configuração do MobDatU-Contact conseguiu superar em média 15% os resultados obtidos pelo MobDatU.

As estratégias de identificação dos contatos mais liberais (*Seguidor-Seguido* e *Retweets* com $\theta = 10\%$) obtiveram resultados piores devido a inferência de contatos ser menos precisa. Ou seja, essas estratégias resultam em considerar contatos que não influenciam a movimentação do usuário, e assim, diminuindo a precisão dos resultados. Com o aumento do valor do θ a definição de contato fica mais restrita o que melhora a precisão do modelo. Porém, para valores muito altos (*e.g.* $\theta = 75\%$) a eficácia do modelo é afetada pelo tamanho pequeno do conjunto de treino.

A Tabela 5.5, mostra o tamanho do conjunto de treino, em número de usuários únicos e número de tuplas $< u_i, r_i, t_i >$. Note que para *Retweets* a medida que o θ aumenta o tamanho do treino diminui. Porém para $\theta = 75\%$ a queda é muito expressiva. Para o Rio de Janeiro o número de usuários cai de 392 ($\theta = 50\%$) para somente 78 ($\theta = 75\%$). Essa queda do número de usuários e tuplas, acaba afetando o treino. Já que com o treino muito pequeno, a capacidade de generalização do modelo é limitada, consequentemente, as previsões serão menos precisas.

A Figura 5.8 mostra as taxas de acerto médias por hora da melhor configuração do MobDatU-Contact e do MobDatU para a coleção do Rio de Janeiro com treino no dia 29/06/2014 e teste no dia 13/07/2014. Assim como observado para o MobDatU,

Tabela 5.5: Análise do Número de Usuários e Tuplas $< u_i, r_i, t_i >$ no Treino para as Estratégias de Identificação de Contato

Treino e Teste em Dias Diferentes					
Estratégia		Rio de Janeiro Treino no Dia 29/06/2014		Belo Horizonte Treino no Dia 03/02/2013	
		# Usuários	# $< u_i, r_i, t_i >$	# Usuários	# $< u_i, r_i, t_i >$
Seguidor-Seguido		943	3.122	938	3.745
Retweets	θ=10%	692	2.001	501	2.145
	θ=25%	400	1.425	517	1.245
	θ=50%	392	1.124	476	1.122
	θ=75%	78	331	77	372
Treino e Teste em Dias Iguais					
Estratégia		Rio de Janeiro Treino no Dia Dia 27/11/2013		Belo Horizonte Treino no Dia 11/09/2013	
		# Usuários	# $< u_i, r_i, t_i >$	# Usuários	# $< u_i, r_i, t_i >$
Seguidor-Seguido		523	2.131	445	1.235
Retweets	θ=10%	497	1.994	401	1.133
	θ=25%	325	842	243	944
	θ=50%	315	830	235	922
	θ=75%	55	231	45	251

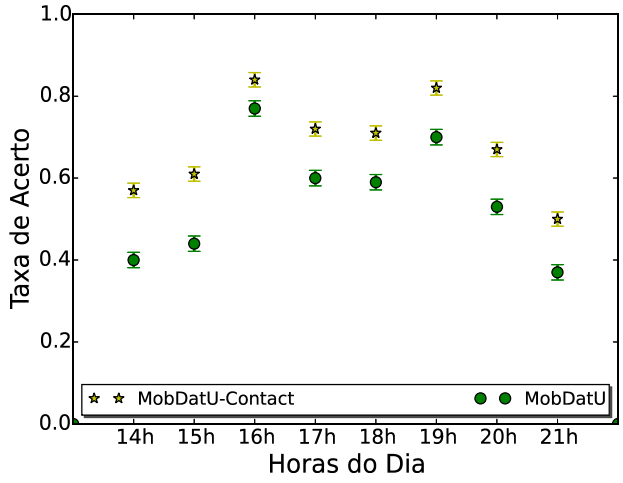


Figura 5.8: Taxas de Acertos Médias por Hora para o MobDatU e MobDatU-Contact na Melhor Configuração - Rio de Janeiro Treino no Dia 29/06/2014 e Teste no Dia 13/07/2014

é possível perceber que quanto maior o volume de dados, maior a taxa de acerto do modelo de previsão. Um exemplo disso é o período de 16h, que apresentou uma maior precisão do que os outros intervalos. Além disso, é possível perceber que os ganhos do MobDatU-Contact sobre MobDatU variaram de 8% a 26%. Esses padrões também foram observados para as demais coleções que pode ser observados nas Figuras A.1(b)-A.19(b) no Apêndice A.

No geral, o MobDatU-Contact conseguiu superar os resultados do outro modelo proposto, o MobDatU, e consequentemente dos dois modelos de referência, SMOOTH e Leap Graph. Ao utilizar a frequência de *retweets*, conseguimos uma melhora nos

resultados da previsão em relação ao MobDatU e aos modelos de referência. Como mostrado, ao considerar as três variáveis (frequência de transição entre regiões, popularidade da região, e probabilidade $P(L_{i,j,t}|C^{\geq n}_{i,j,t})$) para calcular a probabilidade de transição entre as regiões, o resultado da previsão foi melhor do que considerando uma das outras três abordagens. Por último, podemos perceber que uma estratégia muito restrita para identificar os contatos de um usuário (*i.e.*, a estratégia de *Retweets* com limiar $\theta = 75\%$) pode levar a uma redução drástica da quantidade de dados para aprendizado do modelo, o que leva a resultados piores. A perda de precisão do modelo também acontece com estratégias pouco restritas (a estratégia de *Seguidor-Seguido* e *Retweets* com $\theta = 10\%$), pois a inferência dos contatos é menos confiável.

Capítulo 6

Conclusões e Trabalhos Futuros

Nesta dissertação nós propusemos dois novos modelos de previsão de mobilidade humana, o MobDatU e o MobDatU-Contact. Os dois modelos propostos, assim como dois modelos de referência, a saber o SMOOTH e Leap Graph, foram avaliados em diversos cenários, criados a partir de dados coletados de diferentes fontes, utilizados de forma isolada ou combinada.

O MobDatU foi projetado com o objetivo de se adequar bem para diferentes fontes de dados. Para isso, ele utiliza características de modelos de referência que se mostraram eficientes na previsão de mobilidade humana, a saber a popularidade das diferentes regiões e as frequências de transições entre regiões. O MobDatU-Contact, estende o MobDatU para explorar não somente estas duas características mas também os contatos de um usuário na previsão.

Nós avaliamos os dois novos modelos bem como o SMOOTH e o Leap Graph utilizando dados de duas fontes distintas. A primeira fonte foi de uma operadora de telefonia móvel. Esta forneceu dados associados às chamadas realizadas em 5 cidades em 21 dias diferentes, totalizando 27 coleções de dados. A cada chamada está associada uma antena, cuja localização é conhecida e que permite a posição dos usuários quando realiza a chamada. A segunda fonte de dados utilizada foi o Twitter. Para cada cidade e período coberto pelas coleções de dados de telefonia móvel, nós coletamos os *tweets* postados na mesma região e período de tempo. Somente *tweets* com informação geográfica associada foram considerados. Essa informação foi usada para localizar os usuários quando realizaram a postagem de cada *tweet*.

Os resultados experimentais mostraram que nenhum dos dois modelos de referência é superior em todos os cenários investigados, o que demonstra a sensibilidade dos mesmos ao tipo de dados disponíveis. Já o MobDatU se adequou bem a ambos tipos de dados considerados, sendo pelo menos comparável (e por vezes superior) ao melhor

modelo de referência em todos os cenários. O MobDatU conseguiu superar o Leap Graph em até 5% para os cenários de chamada e em 20% para os cenários de *tweets*. Em relação ao SMOOTH, o MobDatU produziu resultados até 120% superiores para os cenários de chamadas e em 15% melhores para os cenários de *tweets*.

O MobDatU-Contact foi avaliado somente para dados do Twitter devido a uma limitação dos nossos dados de telefonia móvel, que não possuem nenhuma informação a partir do qual os contatos de um usuário possam ser inferidos. Os resultados demonstraram que o uso das informações de contato melhoram as previsões, quando comparado ao MobDatU, em até 20% para a melhor estratégia de identificação de contatos de um usuário.

Também foi mostrado que o volume de dados, tanto de chamadas telefônicas quanto de *tweets*, varia bastante durante o dia, o que afeta a precisão de todos os modelos. Ou seja, quanto maior o volume de dados, mais preciso o modelo. Logo, a diferença do volume de dados por período de tempo é algo importante a se considerar quando for realizar a previsão da mobilidade humana.

Uma extensão possível deste trabalho é desenvolver soluções para prever o volume de usuários que estarão em uma determinada região em um certo período. Essa tarefa de previsão, diferente da abordada nesta dissertação, pode ser útil para suportar decisões de gerenciamento e planejamento urbano.

Uma outra direção de trabalho é a investigação de novas estratégias de combinação de dados heterogêneos que permitam a identificação (ou pelo menos a inferência da presença) de um mesmo usuário em múltiplas coleções. Este é um grande desafio técnico uma vez que essas coleções tipicamente são obtidas de forma independente.

Além disso, dividir a área alvo em regiões que possuam aspectos socioculturais similares. Por exemplo, dividir a área em bairros ou de acordo com o volume de pessoas em vez de regiões retangulares de mesmo diâmetro. Isso, pode ajudar a entender melhor os padrões de mobilidade humana em cada região, consequentemente, podendo ajudar no planejamento urbano

O MobDatU-Contact pode ser utilizado com dados de fontes diferentes do Twitter. Como exemplo, dados de chamadas telefônicas que identificam o destinatário de uma ligação, algo que não possuímos nas nossas coleções de dados. A partir desses dados, também é possível investigar diferentes estratégias para inferência dos contatos de um usuário. Está é uma outra direção a ser explorada no futuro.

Referências Bibliográficas

Alharbi, B. & Zhang, X. (2014). Exploring the significance of human mobility patterns in social link prediction. Em *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, SAC '14, pp. 604--609, New York, NY, USA. ACM.

Allamanis, M.; Scellato, S. & Mascolo, C. (2012). Evolution of a location-based online social network: Analysis and models. Em *Proceedings ACM Conference on Internet Measurement*.

Bagrow, J. P.; Wang, D. & Barabási, A.-L. (2011a). Collective response of human populations to large-scale emergencies. *PLoS ONE*, 6(3):e17680.

Bagrow, J. P.; Wang, D. & Barabási, A.-L. (2011b). Collective response of human populations to large-scale emergencies. *PLoS ONE*, 6(3):e17680.

Balcan, D.; Colizza, V.; Gonçalves, B.; Hu, H.; Ramasco, J. J. & Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489.

Becker, R.; Cáceres, R.; Hanson, K.; Isaacman, S.; Loh, J. M.; Martonosi, M.; Rowland, J.; Urbanek, S.; Varshavsky, A. & Volinsky, C. (2013). Human mobility characterization from cellular network data. *Commun. ACM*, 56(1):74--82.

Blondel, V. D.; Decuyper, A. & Krings, G. (2015). A survey of results on mobile phone datasets analysis. *arXiv preprint arXiv:1502.03406*.

Bui, N.; Bui, N.; Michelinakis, F.; Michelinakis, F. & Widmer, J. (2014). A model for throughput prediction for mobile users. Em *Proceedings 20th European Wireless Conference;*.

Calabrese, F.; Pereira, F.; Di Lorenzo, G.; Liu, L. & Ratti, C. (2010). The geography of taste: Analyzing cell-phone mobility and social events. Em Floréen, P.; Krüger,

A. & Spasojevic, M., editores, *Pervasive Computing*, volume 6030 of *Lecture Notes in Computer Science*, pp. 22–37. Springer Berlin / Heidelberg.

Candia, J.; González, M. C.; Wang, P.; Schoenharl, T.; Madey, G. & Barabási, A.-L. (2008). Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015.

Cheng, Z.; Caverlee, J.; Lee, K. & Sui, D. Z. (2011). Exploring millions of footprints in location sharing services. Em *In ICWSM 2011*.

Chiu, R.-F.; Yeh, Y.-S.; Chi, S.; Lee, R.; Wu, A.; Chang, H.-J. & Chang, L. (2009). Kaohsiung county broadband mobile network. *Software Engineering, Artificial Intelligences, Networking and Parallel/Distributed Computing, 2009. SNPD '09. 10th ACIS International Conference on*.

Cho, E.; Myers, S. A. & Leskovec, J. (2011). Friendship and mobility: User movement in location-based social networks. Em *Proc. 17th ACM International Conference on Knowledge Discovery and Data Mining*.

Csáji, B. C.; Browet, A.; Traag, V. A.; Delvenne, J.-C.; Huens, E.; Van Dooren, P.; Smoreda, Z. & Blondel, V. D. (2013). Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and its Applications*, 392(6):1459--1473.

Davis Jr., C. A.; Pappa, G. L.; de Oliveira, D. R. R. & de L. Arcanjo, F. (2011). Inferring the location of twitter messages based on user relationships. *Transactions in GIS*, 15(6):735--751.

Dong, W.; Duffield, N.; Ge, Z.; Lee, S. & Pang, J. (2013). Modeling cellular user mobility using a leap graph. Em *Proc. 14th International Conference on Passive and Active Measurement*.

Eagle, N.; Pentland, A. & Lazer, D. (2009). Inferring social network structure using mobile phone data. *Proceedings of the National Academy of Sciences (PNAS)*, 106(36):15274–15278.

Erman, J. & Ramakrishnan, K. (2013). Understanding the super-sized traffic of the super bowl. Em *Proceedings of the 2013 Conference on Internet Measurement Conference*, IMC '13, pp. 353--360, New York, NY, USA. ACM.

Gonzalez, M. (2013). Unraveling daily human mobility motifs. Em *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, pp. xxxviii–xxxviii.

González, M. C.; Hidalgo, C. A. & Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature Publishing Group*, 453.

Herrmann, K. (2003). Modeling the sociological aspects of mobility in ad hoc networks. Em *Proceedings of the 6th ACM International Workshop on Modeling Analysis and Simulation of Wireless and Mobile Systems*, MSWIM '03, pp. 128--129, New York, NY, USA. ACM.

Isaacman, S.; Becker, R.; Cáceres, R.; Martonosi, M.; Rowland, J.; Varshavsky, A. & Willinger, W. (2012). Human mobility modeling at metropolitan scales. Em *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, MobiSys '12, pp. 239--252, New York, NY, USA. ACM.

Jeung, H.; Liu, Q.; Shen, H. T. & Zhou, X. (2008). A hybrid prediction model for moving objects. Em *Proc. IEEE 24th International Conference on Data Engineering*.

Jia, Y.; Wang, Y.; Jin, X. & Cheng, X. (2014). Tsbm: The temporal-spatial bayesian model for location prediction in social networks. Em *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*, volume 2, pp. 194–201.

Jiang, S.; Ferreira, J. & González, M. (2012). Clustering daily patterns of human activities in the city. *Data Mining and Knowledge Discovery*, 25(3):478–510.

Kim, M.; Kotz, D. & Kim, S. (2006). Extracting a mobility model from real user traces. Em *INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings*, pp. 1–13.

Klebanov, L. B. (2003). *Heavy tailed distributions*. Matfyzpress.

Koo, J. & Chung, K. (2010). Marc: Adaptive rate control scheme for improving the qoe of streaming services in mobile broadband networks. *Communications and Information Technologies (ISCIT), 2010 International Symposium on*.

Lee, K.; Hong, S.; Kim, S. J.; Rhee, I. & Chong, S. (2009). Slaw: A new mobility model for human walks. Em *Proc. INFOCOM*.

Lee, K.; Hong, S.; Kim, S. J.; Rhee, I. & Chong, S. (2012). Slaw: Self-similar least-action human walk. *IEEE/ACM Transactions on Networking*, 20(2):515–529.

Li, N. & Chen, G. (2009a). Analysis of a location-based social network. Em *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04*, CSE '09, pp. 263--270, Washington, DC, USA. IEEE Computer Society.

Li, N. & Chen, G. (2009b). Multi-layered friendship modeling for location-based mobile social networks. Em *Mobile and Ubiquitous Systems: Networking Services, MobiQuitous, 2009. MobiQuitous '09. 6th Annual International*, pp. 1–10.

Liu, S.; Liu, Y.; Ni, L. M.; Fan, J. & Li, M. (2010). Towards mobility-based clustering. Em *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pp. 919--928, New York, NY, USA. ACM.

Munjal, A.; Camp, T. & Navidi, W. C. (2011). Smooth: A simple way to model human mobility. Em *Proc. 14th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*.

Musolesi, M. & Mascolo, C. (2007). Designing mobility models based on social network theory. *SIGMOBILE Mobile Computing Communications Review*, 11(3).

Musolesi, M. & Mascolo, C. (2008). Mobility models for systems evaluation – a survey.

Navidi, W.; Camp, T. & Bauer, N. (2004). Improving the accuracy of random waypoint simulations through steady-state initialization.

Nguyen, T. & Szymanski, B. K. (2012). Using location-based social networks to validate human mobility and relationships models. Em *Proc. International Conference on Advances in Social Networks Analysis and Mining*.

Noulas, A. & Mascolo, C. (2013). Exploiting foursquare and cellular data to infer user activity in urban environments. Em *Mobile Data Management (MDM), 2013 IEEE 14th International Conference on*, volume 1, pp. 167–176.

Noulas, A.; Scellato, S.; Lambiotte, R.; Pontil, M. & Mascolo, C. (2012a). A tale of many cities: universal patterns in human urban mobility. *PloS one*, 7(5):e37027.

Noulas, A.; Scellato, S.; Lathia, N. & Mascolo, C. (2012b). A random walk around the city: New venue recommendation in location-based social networks. Em *Proc. International Conference on Social Computing*.

Rhee, I.; Shin, M.; Hong, S.; Lee, K. & Chong, S. (2008). On the levy-walk nature of human mobility. Em *Proc. INFOCOM*.

Scellato, S.; Noulas, A.; Lambiotte, R. & Mascolo, C. (2011). Socio-spatial properties of online location-based social networks. Em Adamic, L. A.; Baeza-Yates, R. A. & Counts, S., editores, *ICWSM*. The AAAI Press.

Silva, T.; de Melo, P.; Viana, A.; Almeida, J.; Salles, J. & Loureiro, A. (2013). Traffic condition is more than colored lines on a map: Characterization of waze alerts. Em Jatowt, A.; Lim, E.-P.; Ding, Y.; Miura, A.; Tezuka, T.; Dias, G.; Tanaka, K.; Flanagin, A. & Dai, B., editores, *Social Informatics*, volume 8238 of *Lecture Notes in Computer Science*, pp. 309–318. Springer International Publishing.

Song, C.; Koren, T.; Wang, P. & Barabási, A.-L. (2010). Modelling the scaling properties of human mobility. *Nature Publishing Group*, 6.

Soper, D. (2012). Is human mobility tracking a good idea? *Communications of the ACM*, 55(4):35--37.

Tasse, F. P. & Glass, K. (2008). *Crowd simulation of pedestrians in a virtual city.* Tese de doutorado, Masters thesis, Rhodes University.

Tostes, A. I. J.; de L. P. Duarte-Figueiredo, F.; Assunção, R.; Salles, J. & Loureiro, A. A. F. (2013). From data to knowledge: City-wide traffic flows analysis and prediction using bing maps. Em *Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing*, UrbComp '13, pp. 12:1--12:8, New York, NY, USA. ACM.

Urgaonkar, R. & Neely, M. (2011). Network capacity region and minimum energy function for a delay-tolerant mobile ad hoc network. *Networking, IEEE/ACM Transactions on*, 19(4):1137 –1150.

Vespignani, A. (2009). Predicting the behavior of techno-social systems. *Science*, 325(5939):425–428.

Walsh, C.; Doci, A. & Camp, T. (2008). A call to arms: It's time for real mobility models. *SIGMOBILE Mob. Comput. Commun. Rev.*, 12(1):34--36.

Wang, D.; Pedreschi, D.; Song, C.; Giannotti, F. & Barabasi, A.-L. (2011). Human mobility, social ties, and link prediction. Em *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pp. 1100--1108, New York, NY, USA. ACM.

Wang, D. & Song, C. (2015). Impact of human mobility on social networks. *Communications and Networks, Journal of*, 17(2):100–109.

Wang, P.; González, M. C.; Hidalgo, C. A. & Barabási, A.-L. (2009). Understanding the spreading patterns of mobile phone viruses. *Science*, 324(5930):1071–1076.

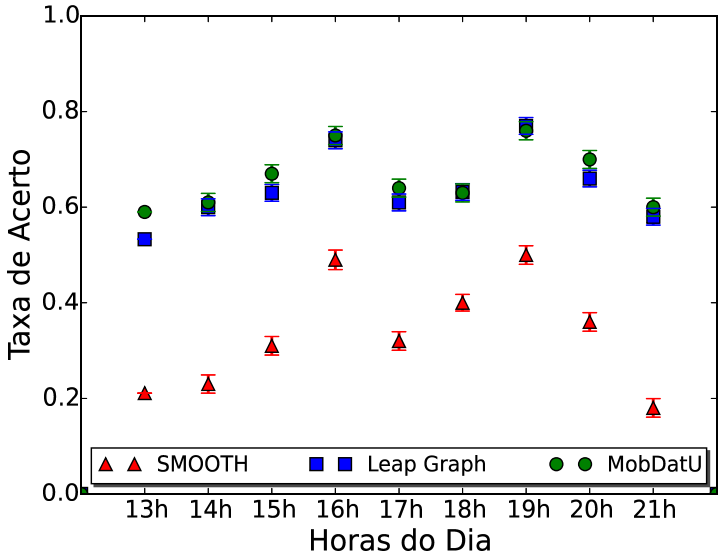
Xavier, F. H. Z.; Silveira, L. M.; Almeida, J. M. d.; Ziviani, A.; Malab, C. H. S. & Marques-Neto, H. T. (2012). Analyzing the workload dynamics of a mobile phone network in large scale events. Em *Proc. First Workshop on Urban Networking*.

Zheng, Y. & Xie, X. (2011). Learning travel recommendations from user-generated gps traces. *ACM Transactions on Intelligent Systems Technology*, 2(1).

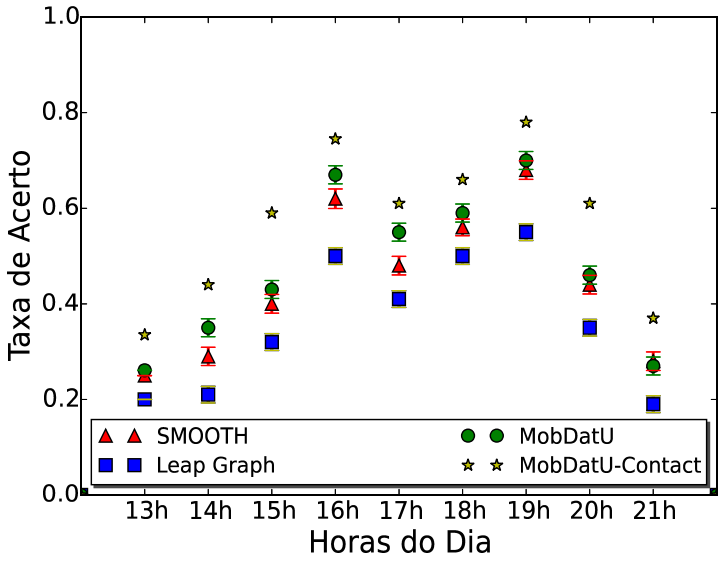
Apêndice A

APÊNDICE A - Avaliação dos Modelos para Janelas de Tempo em Todas as Coleções

As Figuras A.1 - A.19 mostram as taxas de acerto médias por período de hora de todos os modelos para todas as coleções nos cenários de dados homogêneos. Já as Figuras A.20 - A.38 mostram as taxas de acerto médias por período de hora de todos os modelos para todas as coleções nos cenários de dados heterogêneos. Em todos os casos quanto maior o volume de dados do período de tempo, maior a taxa de acerto média dos dois modelos propostos e dos dois modelos de referência, conforme discutido no Capítulo 5.

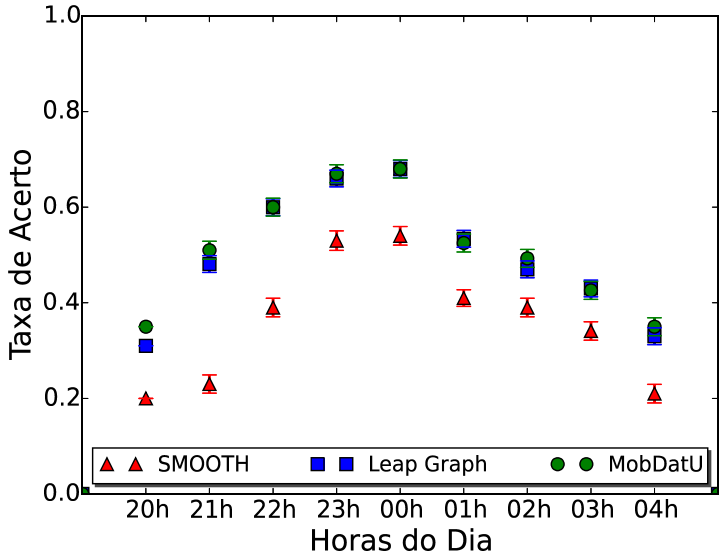


(a) Chamadas

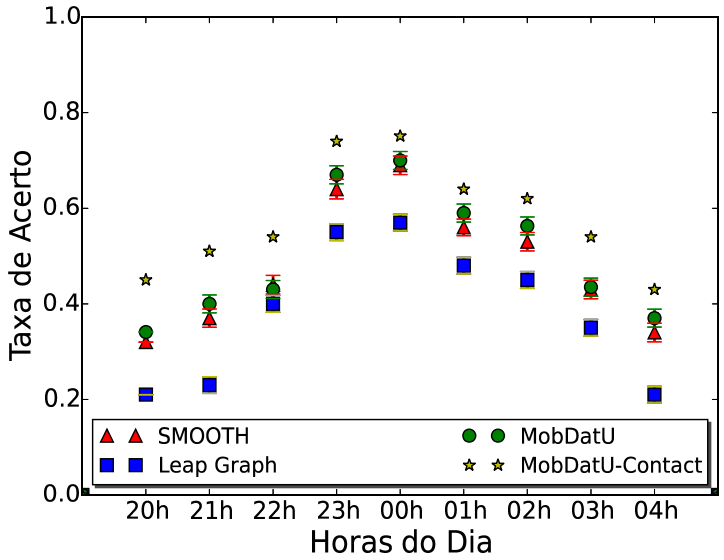


(b) *Tweets*

Figura A.1: Taxas de Acerto Médias por Hora para os Cenário de Dados Homogêneos - Belo Horizonte Treino e Teste no Dia 21/10/2011

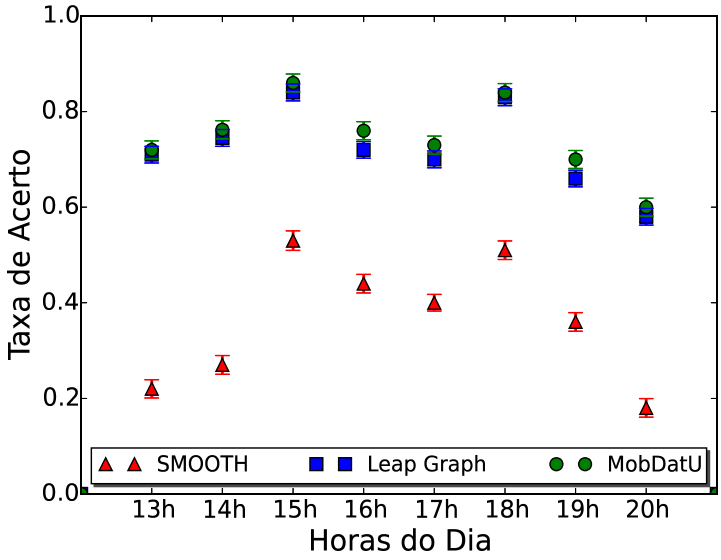


(a) Chamadas

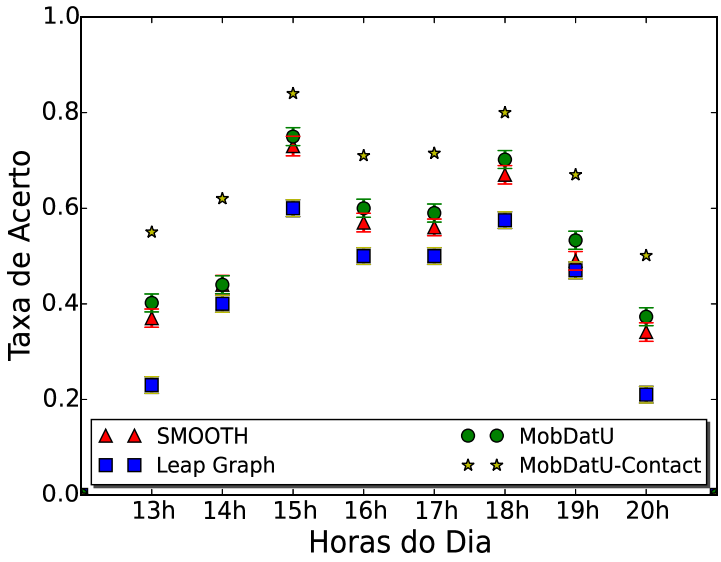


(b) *Tweets*

Figura A.2: Taxas de Acerto Médias por Hora para os Cenário de Dados Homogêneos - Belo Horizonte Treino no Dia 31/12/2011 e Teste no Dia 03/01/2012

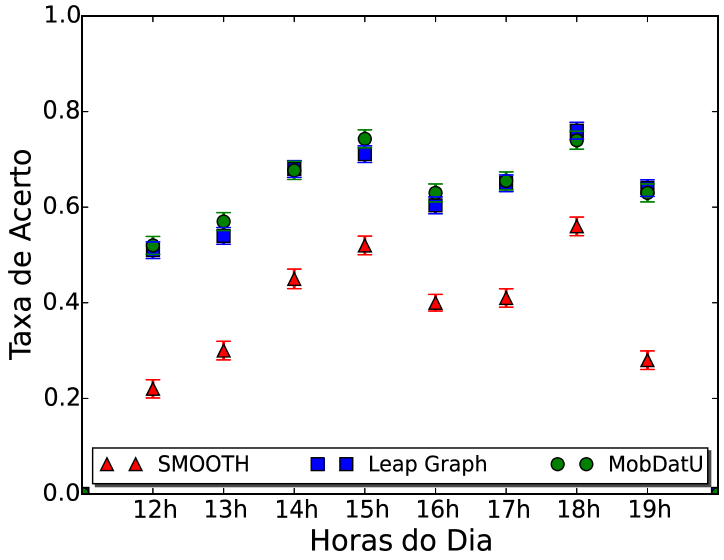


(a) Chamadas

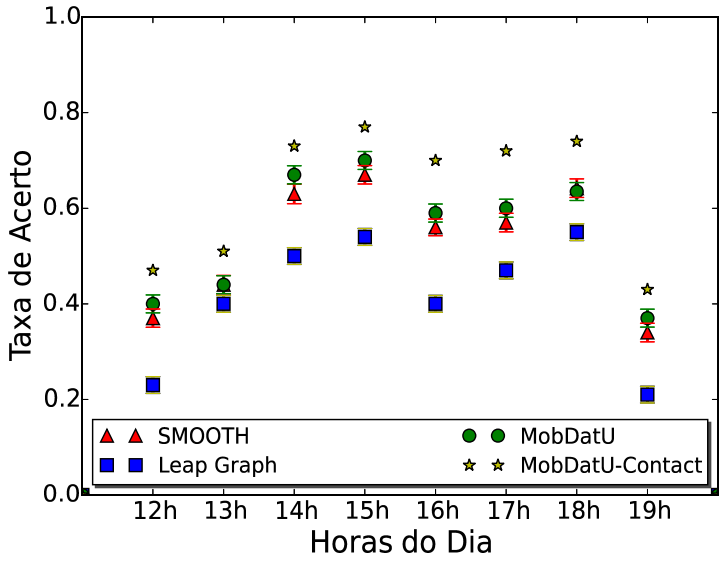


(b) *Tweets*

Figura A.3: Taxas de Acerto Médias por Hora para os Cenário de Dados Homogêneos - Belo Horizonte Treino no Dia 03/02/2013 e Teste no Dia 10/03/2013

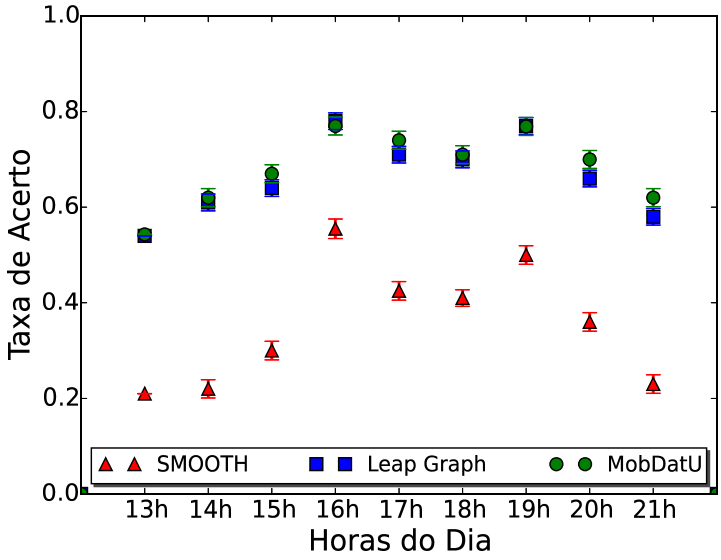


(a) Chamadas

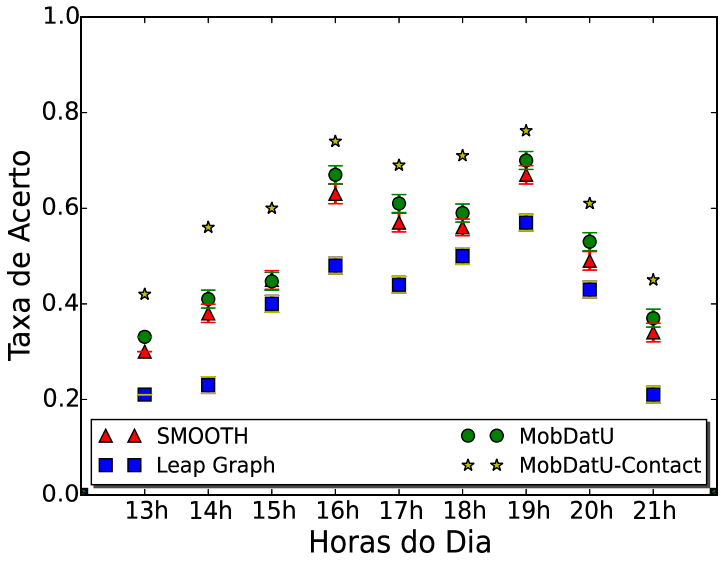


(b) *Tweets*

Figura A.4: Taxas de Acerto Médias por Hora para os Cenário de Dados Homogêneos - Belo Horizonte Treino e Teste no Dia 02/03/2013

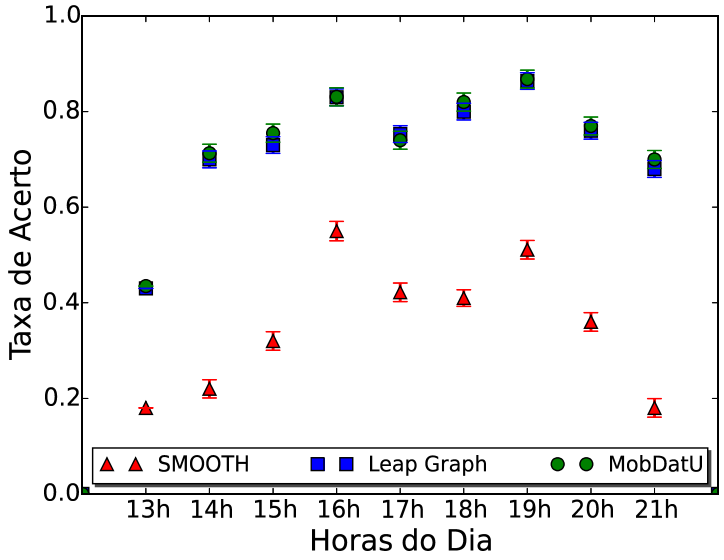


(a) Chamadas

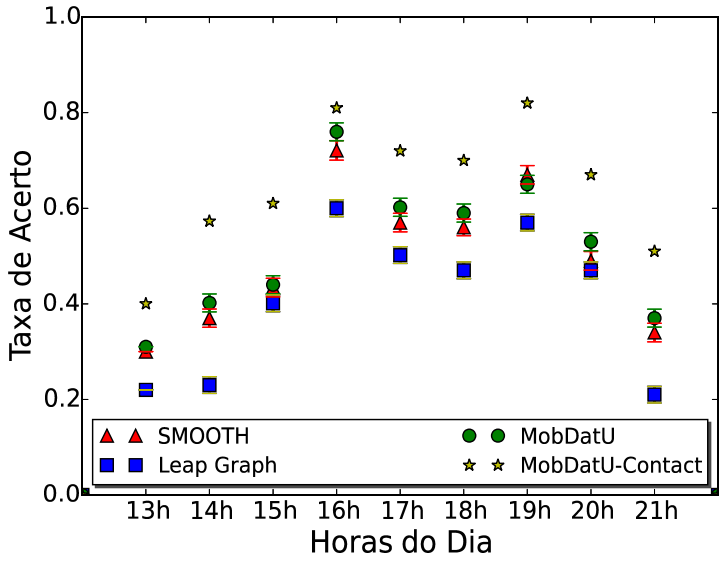


(b) *Tweets*

Figura A.5: Taxas de Acerto Médias por Hora para os Cenário de Dados Homogêneos - Belo Horizonte Treino e Teste no Dia 22/06/2013

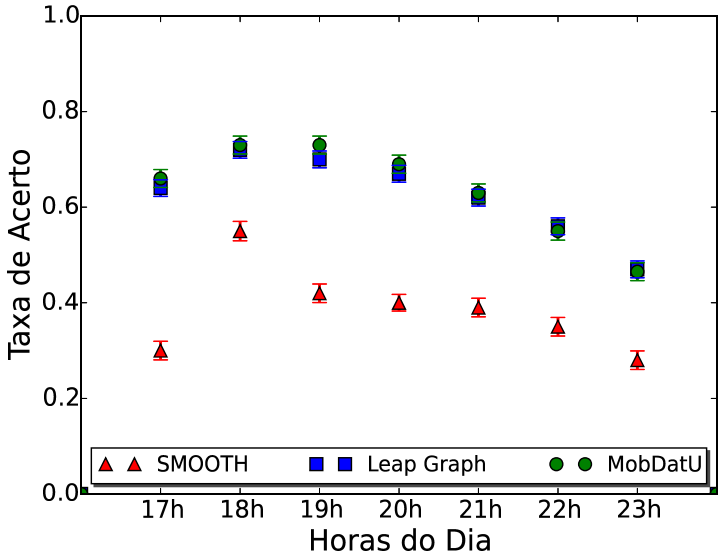


(a) Chamadas

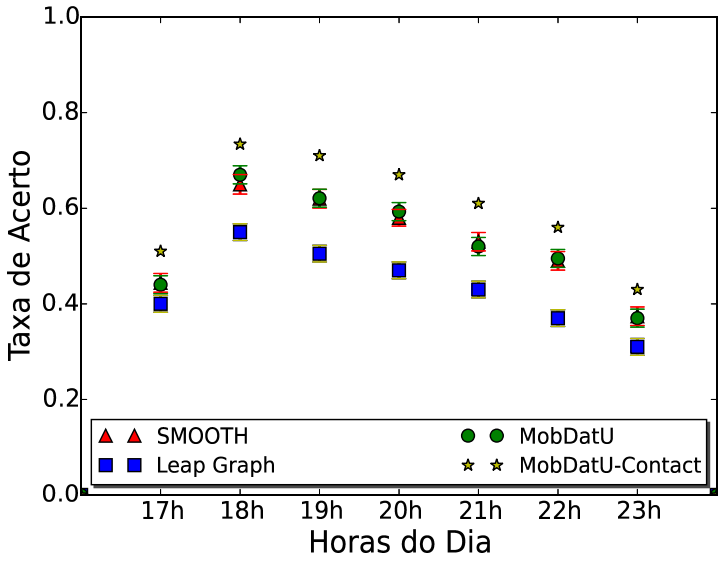


(b) *Tweets*

Figura A.6: Taxas de Acerto Médias por Hora para os Cenário de Dados Homogêneos - Belo Horizonte Treino e Teste no Dia 26/06/2013

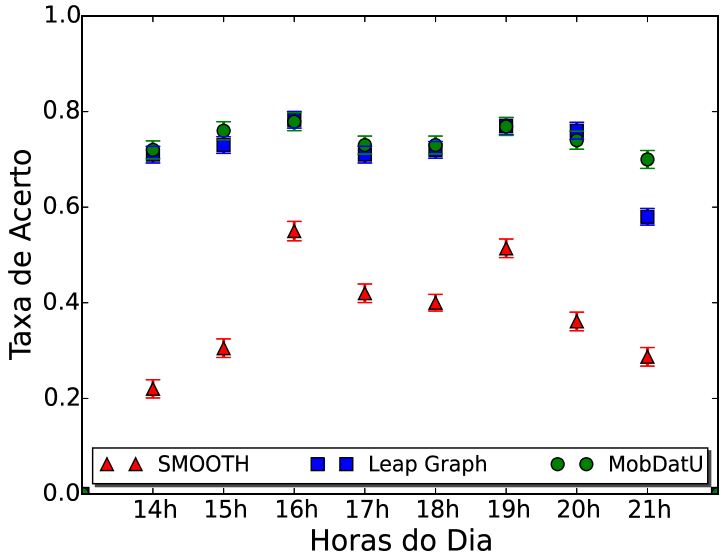


(a) Chamadas

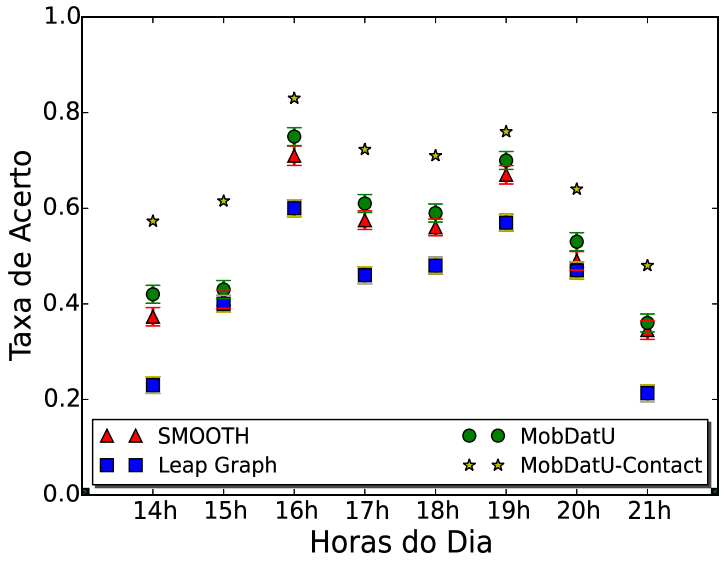


(b) *Tweets*

Figura A.7: Taxas de Acerto Médias por Hora para os Cenário de Dados Homogêneos - Belo Horizonte Treino e Teste no Dia 11/09/2013

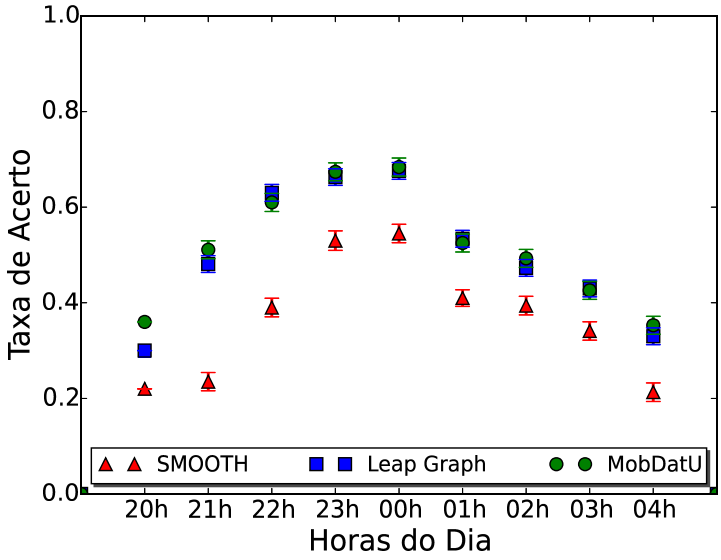


(a) Chamadas

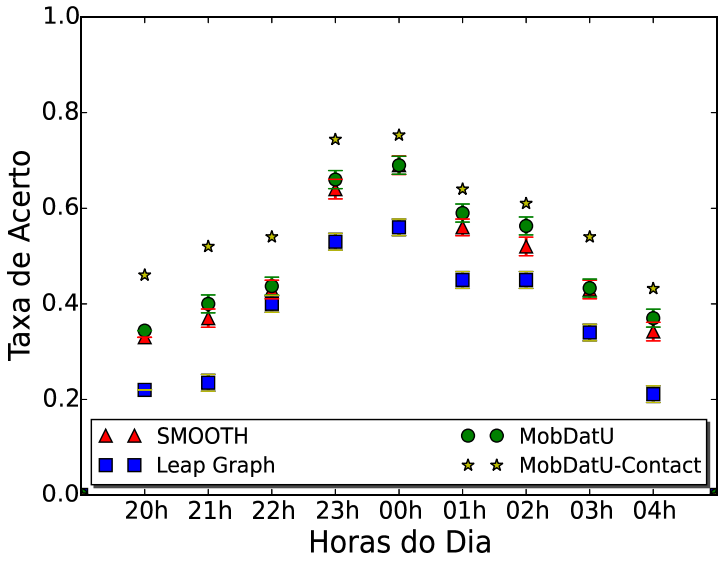


(b) *Tweets*

Figura A.8: Taxas de Acerto Médias por Hora para os Cenário de Dados Homogêneos - Fortaleza Treino e Teste no Dia 29/06/2014

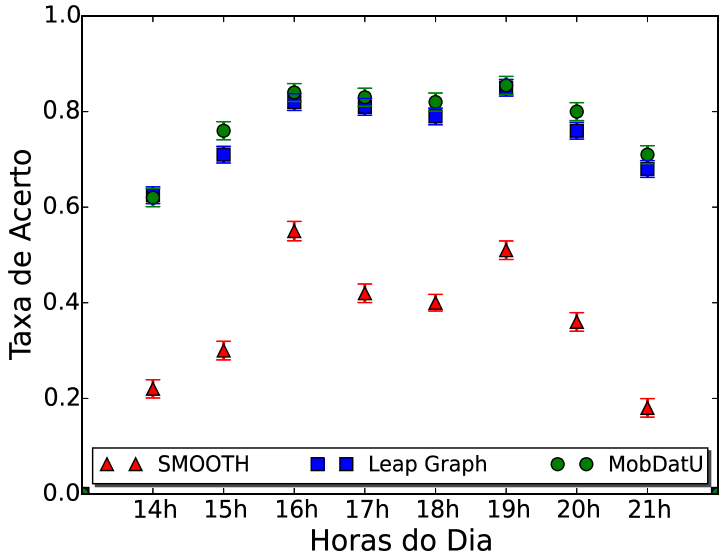


(a) Chamadas

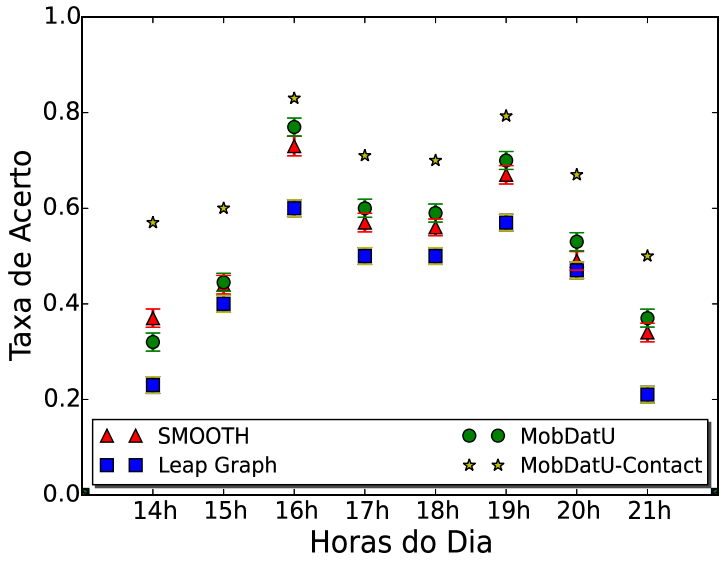


(b) *Tweets*

Figura A.9: Taxas de Acerto Médias por Hora para os Cenário de Dados Homogêneos - Recife Treino no Dia 31/12/2011 e Teste no Dia 03/01/2012

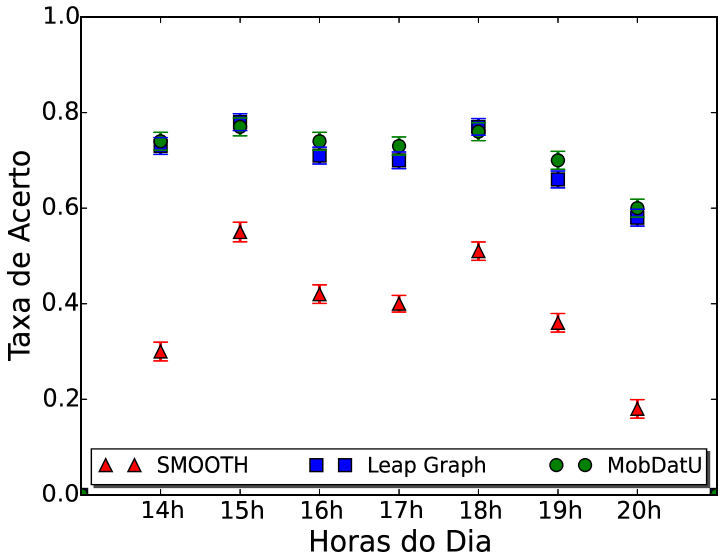


(a) Chamadas

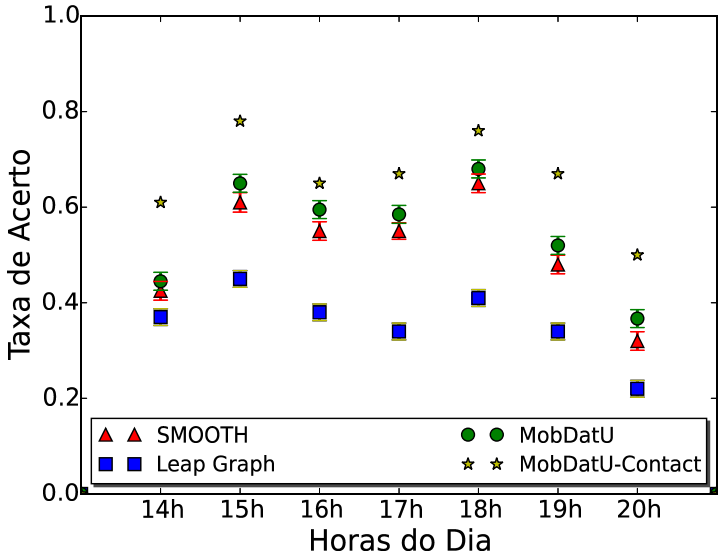


(b) *Tweets*

Figura A.10: Taxas de Acerto Médias por Hora para os Cenário de Dados Homogêneos - Recife Treino e Teste no Dia 29/06/2014

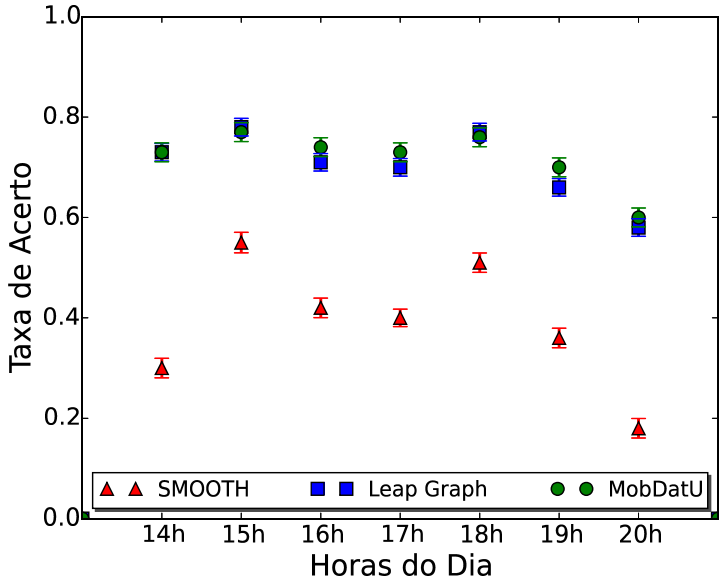


(a) Chamadas

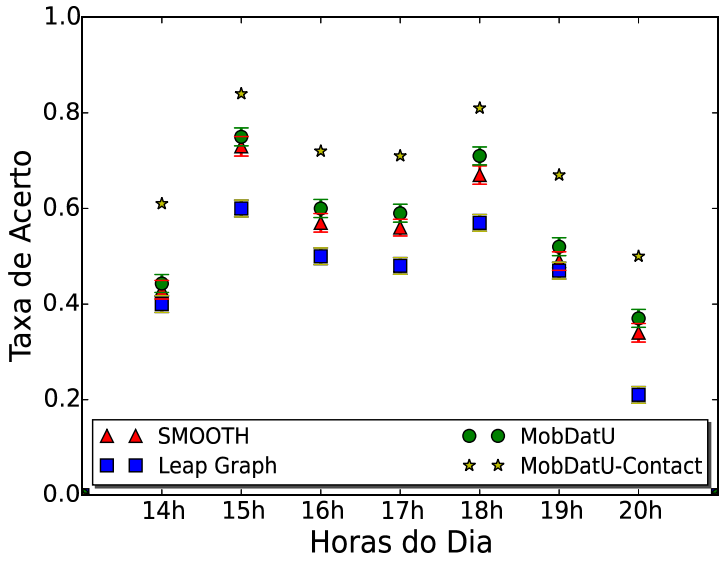


(b) *Tweets*

Figura A.11: Taxas de Acerto Médias por Hora para os Cenário de Dados Homogêneos - Rio de Janeiro Treino Dia 28/08/2011 e Teste no Dia 30/10/2011

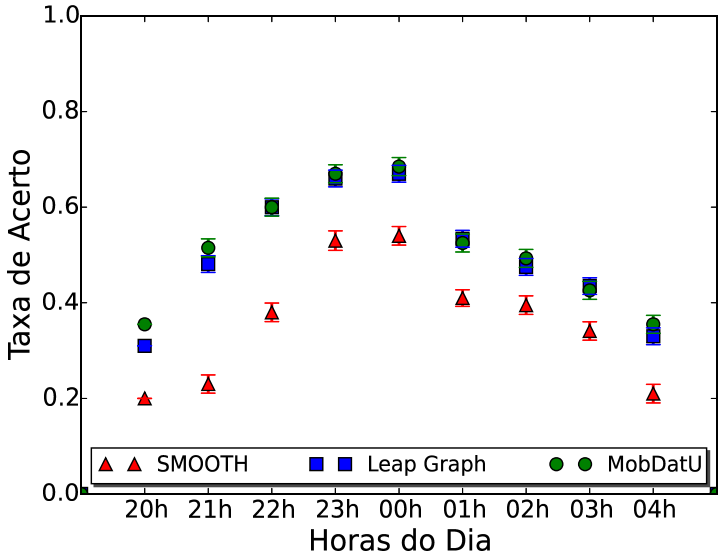


(a) Chamadas

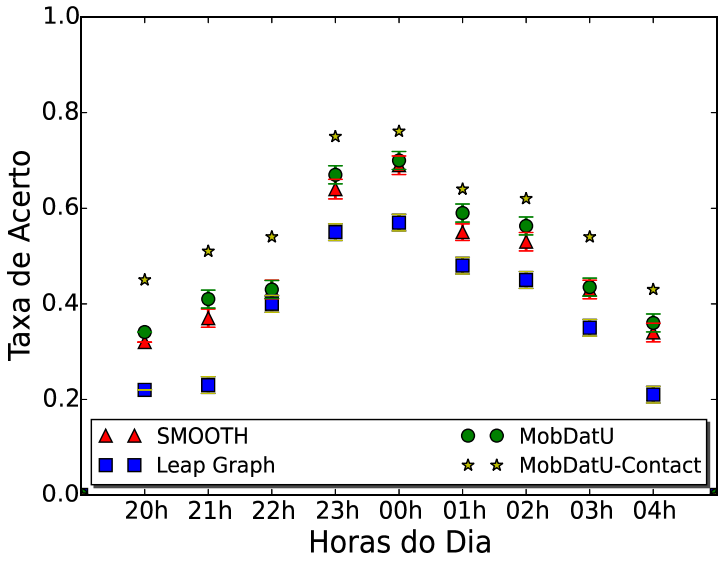


(b) *Tweets*

Figura A.12: Taxas de Acerto Médias por Hora para os Cenário de Dados Homogêneos - Rio de Janeiro Treino Dia 04/12/2011 e Teste no Dia 11/12/2011

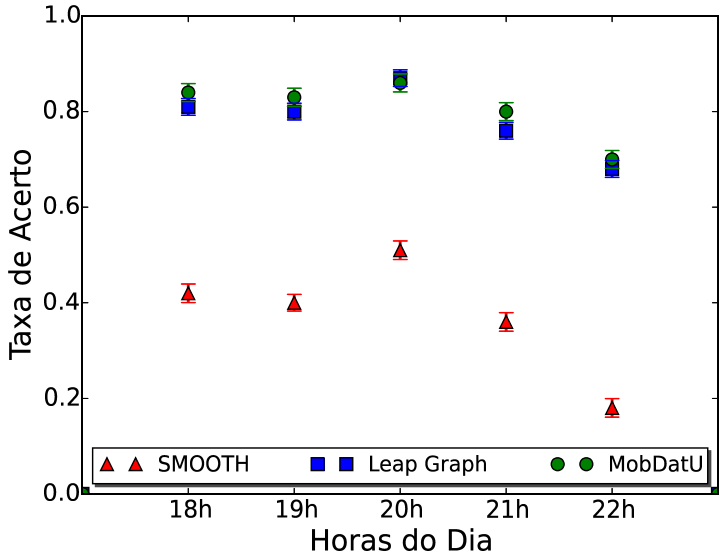


(a) Chamadas

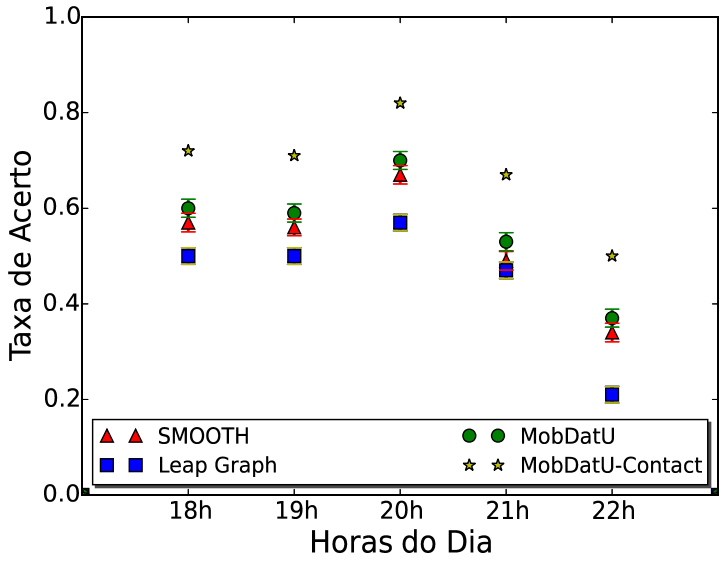


(b) *Tweets*

Figura A.13: Taxas de Acerto Médias por Hora para os Cenário de Dados Homogêneos - Rio de Janeiro Treino Dia 31/12/2011 e Teste no Dia 03/01/2012

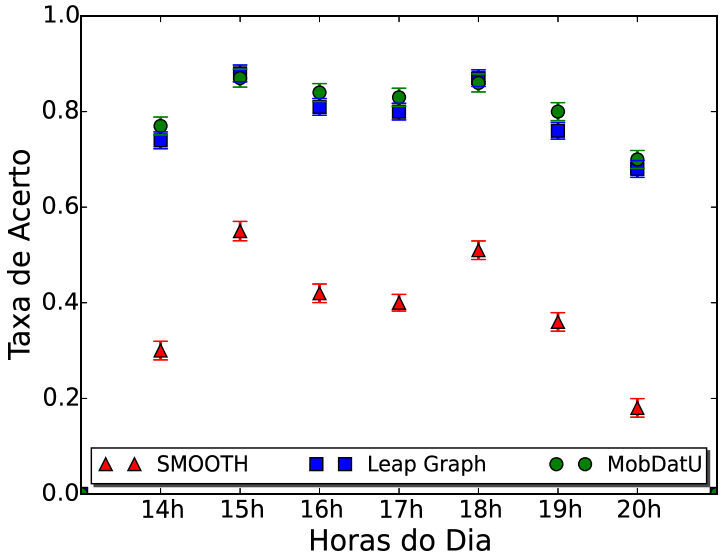


(a) Chamadas

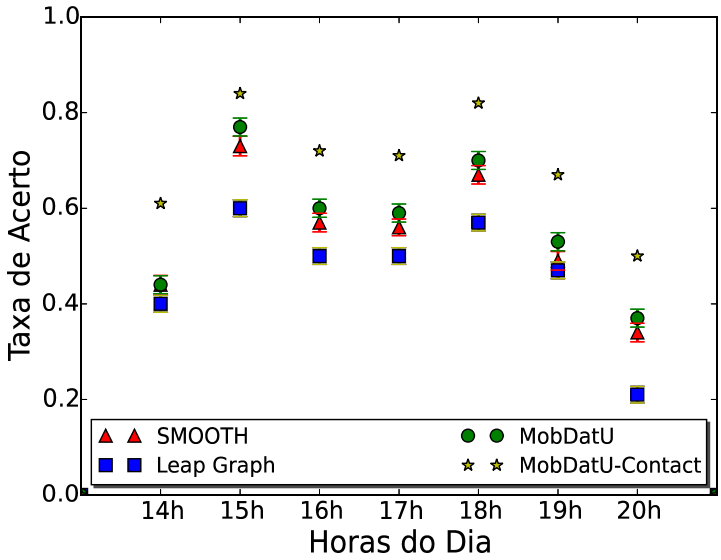


(b) *Tweets*

Figura A.14: Taxas de Acerto Médias por Hora para os Cenário de Dados Homogêneos - Rio de Janeiro Treino e Teste no Dia 29/03/2012

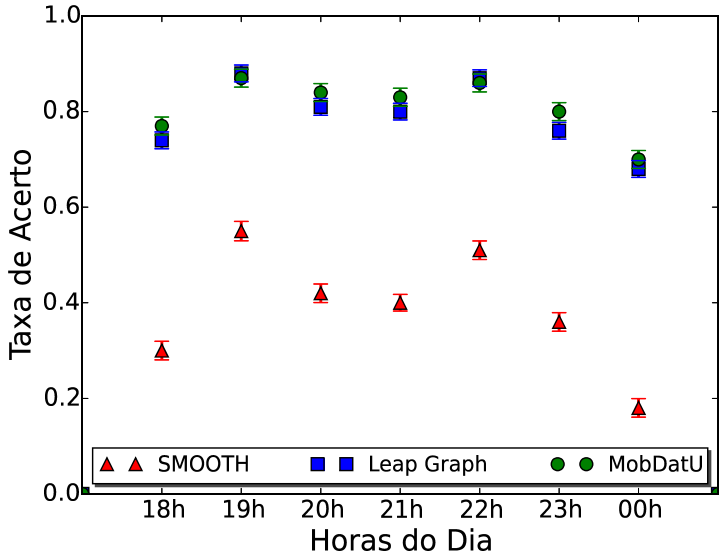


(a) Chamadas

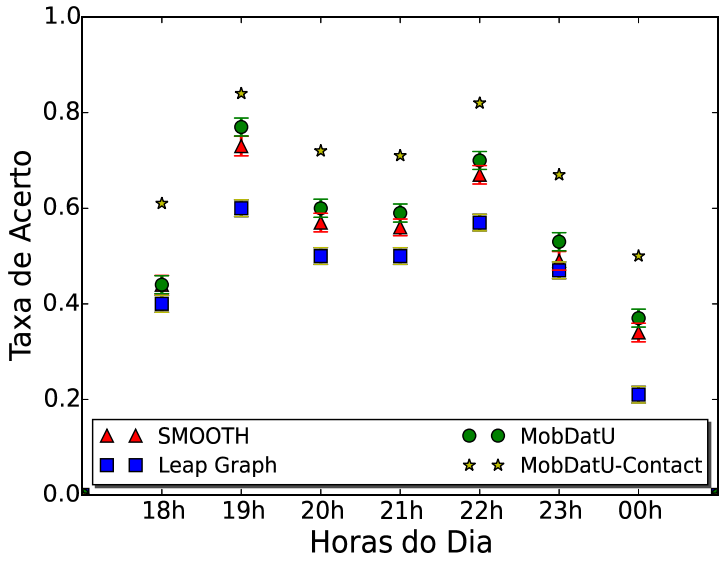


(b) *Tweets*

Figura A.15: Taxas de Acerto Médias por Hora para os Cenário de Dados Homogêneos - Rio de Janeiro Treino e Teste no Dia 08/07/2012

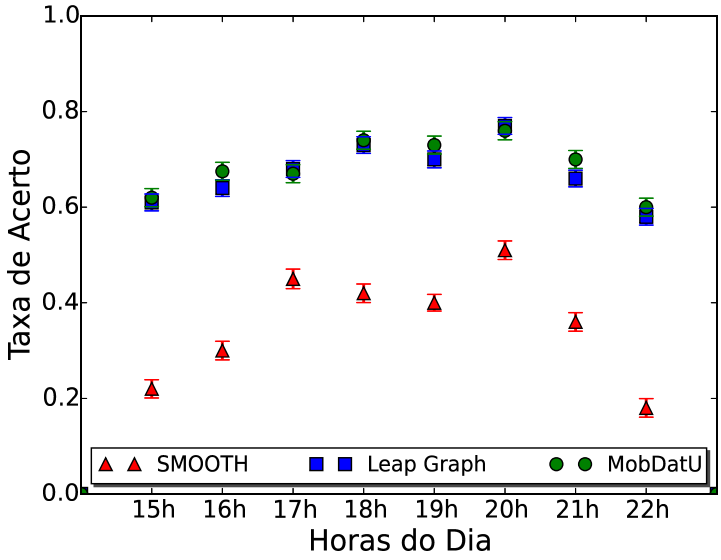


(a) Chamadas

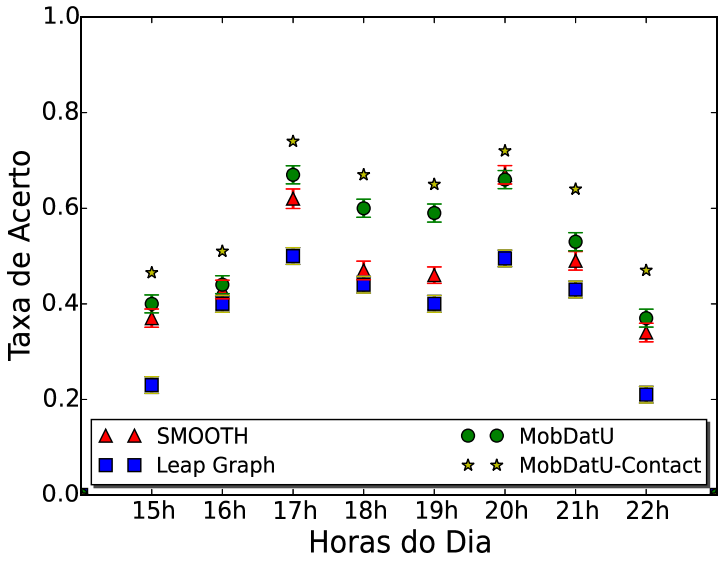


(b) *Tweets*

Figura A.16: Taxas de Acerto Médias por Hora para os Cenário de Dados Homogêneos - Rio de Janeiro Treino e Teste no Dia 27/11/2013

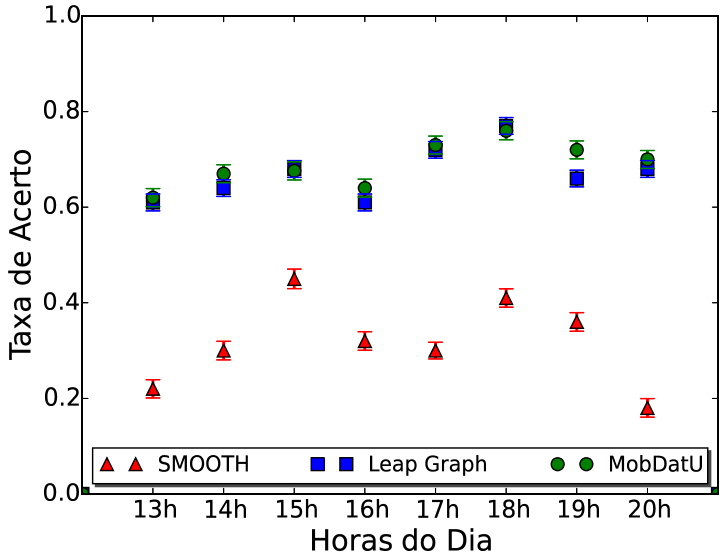


(a) Chamadas

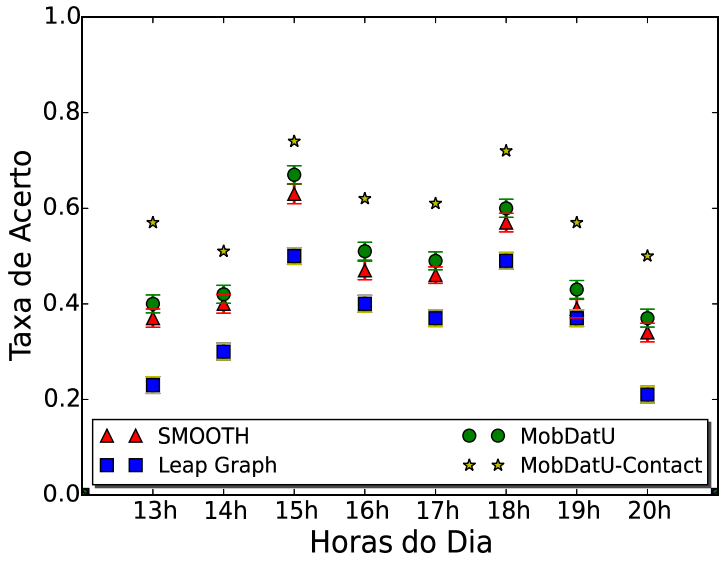


(b) *Tweets*

Figura A.17: Taxas de Acerto Médias por Hora para os Cenário de Dados Homogêneos - São Paulo Treino e Teste no Dia 04/02/2012

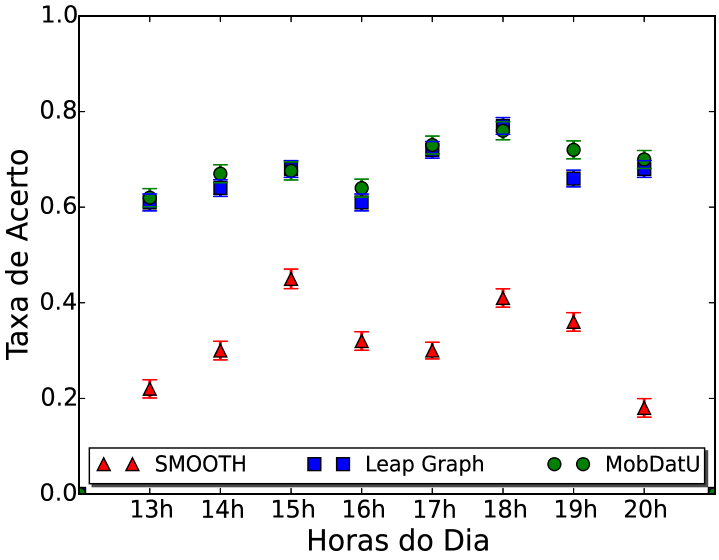


(a) Chamadas

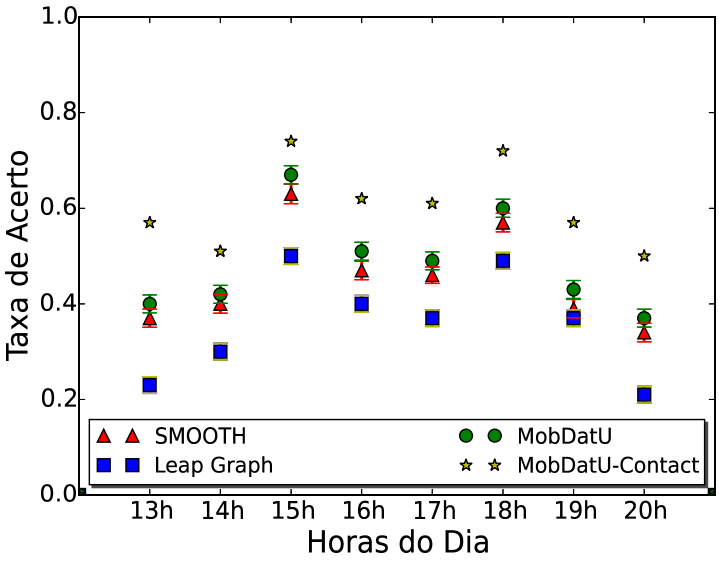


(b) *Tweets*

Figura A.18: Taxas de Acerto Médias por Hora para os Cenário de Dados Homogêneos - São Paulo Treino e Teste no Dia 25/11/2012

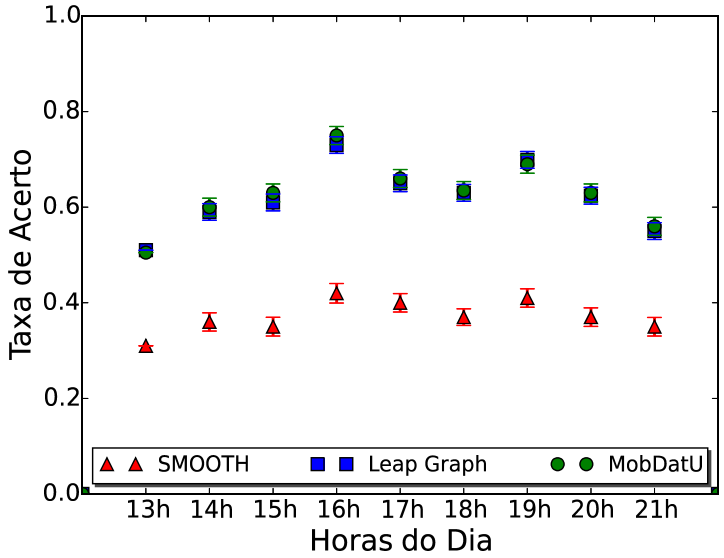


(a) Chamadas

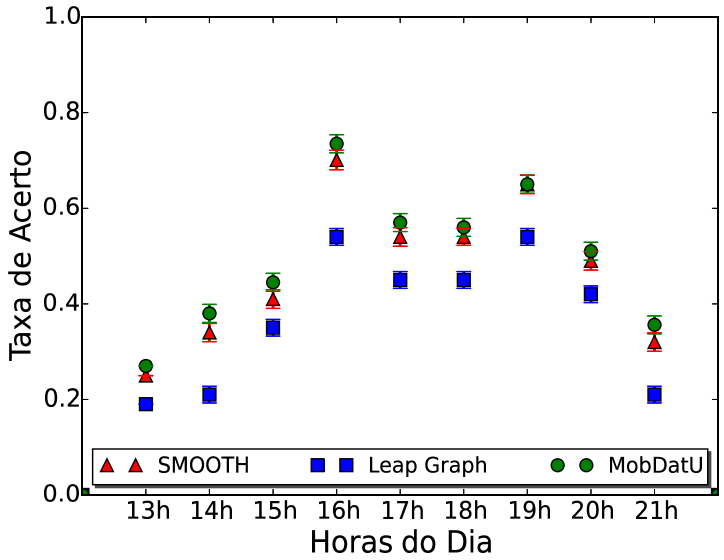


(b) *Tweets*

Figura A.19: Taxas de Acerto Médias por Hora para os Cenário de Dados Homogêneos - São Paulo Treino e Teste no Dia 24/03/2013

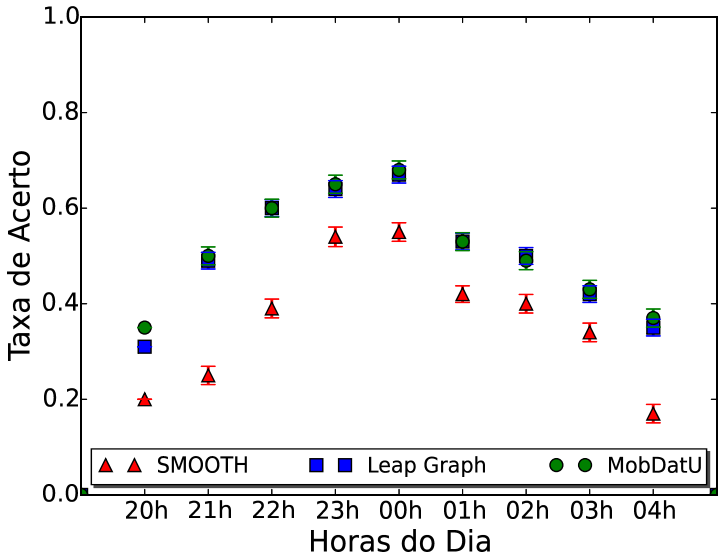


(a) *Tweets* para Chamadas

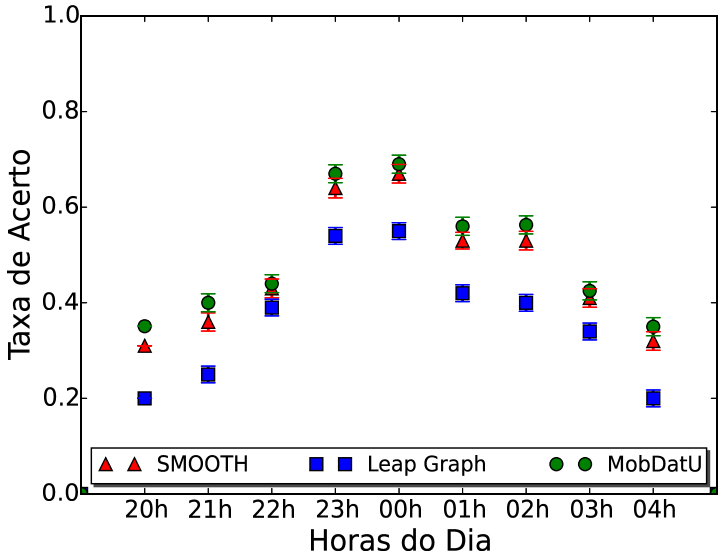


(b) Chamadas para *Tweets*

Figura A.20: Taxas de Acerto Médias por Hora para os Cenário de Dados Heterogêneos - Belo Horizonte Treino e Teste no Dia 21/10/2011

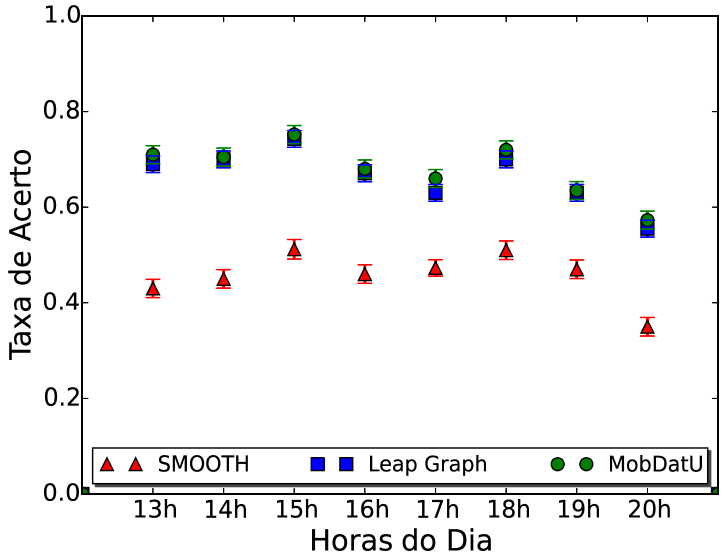


(a) *Tweets* para Chamadas

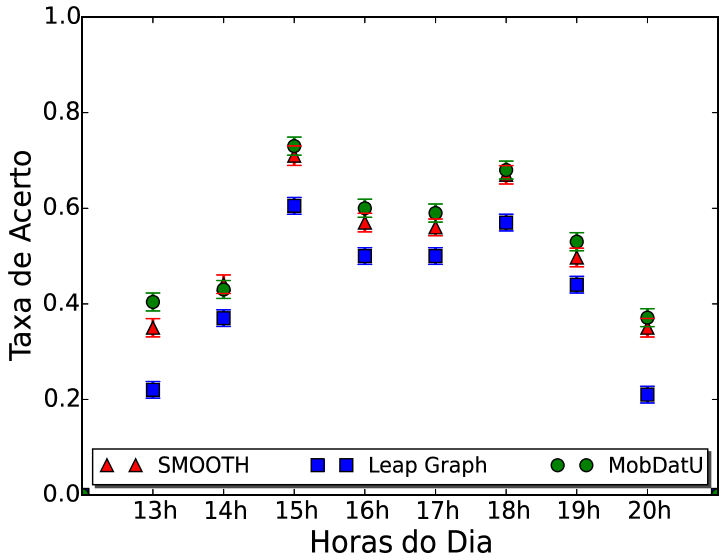


(b) Chamadas para *Tweets*

Figura A.21: Taxas de Acerto Médias por Hora para os Cenário de Dados Heterogêneos - Belo Horizonte Treino no Dia 31/12/2011 e Teste no Dia 03/01/2012

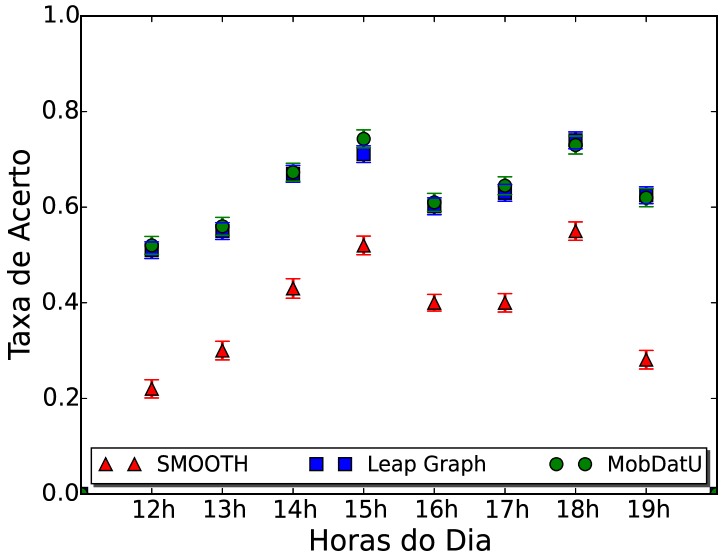


(a) *Tweets* para Chamadas

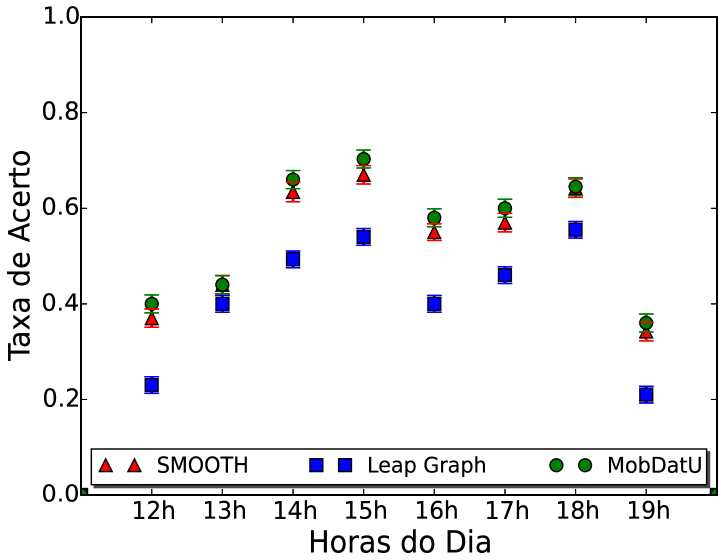


(b) Chamadas para *Tweets*

Figura A.22: Taxas de Acerto Médias por Hora para os Cenário de Dados Heterogêneos - Belo Horizonte Treino no Dia 03/02/2013 e Teste no Dia 10/03/2013

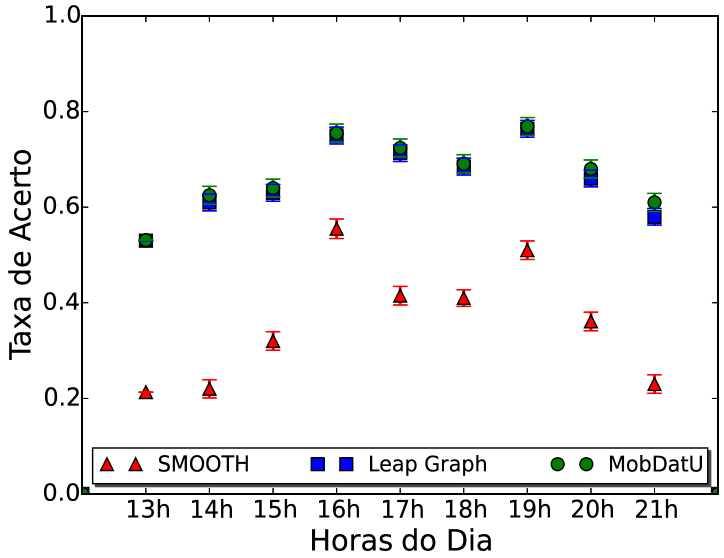


(a) *Tweets* para Chamadas

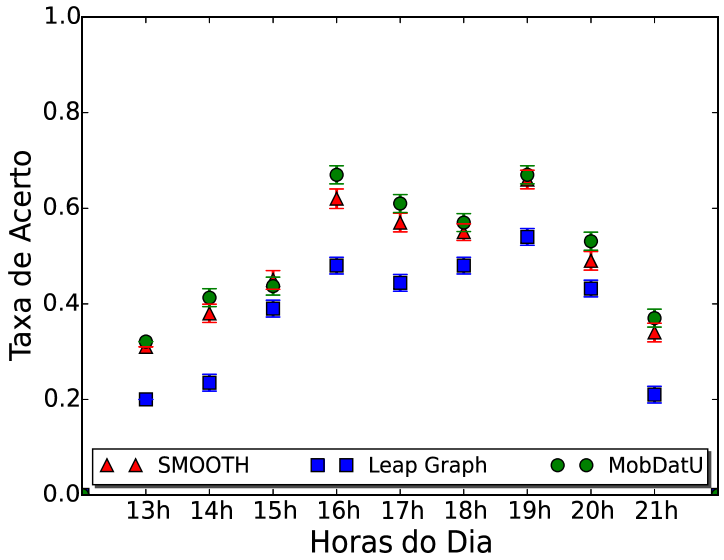


(b) Chamadas para *Tweets*

Figura A.23: Taxas de Acerto Médias por Hora para os Cenário de Dados Heterogêneos - Belo Horizonte Treino e Teste no Dia 02/03/2013

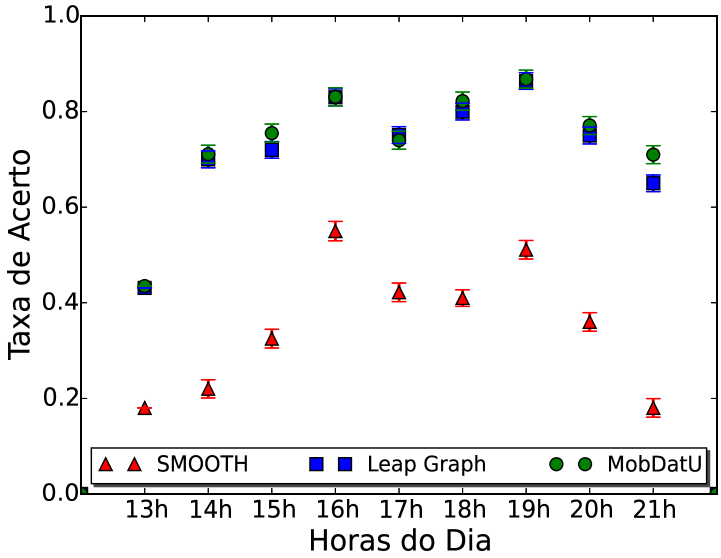


(a) *Tweets* para Chamadas

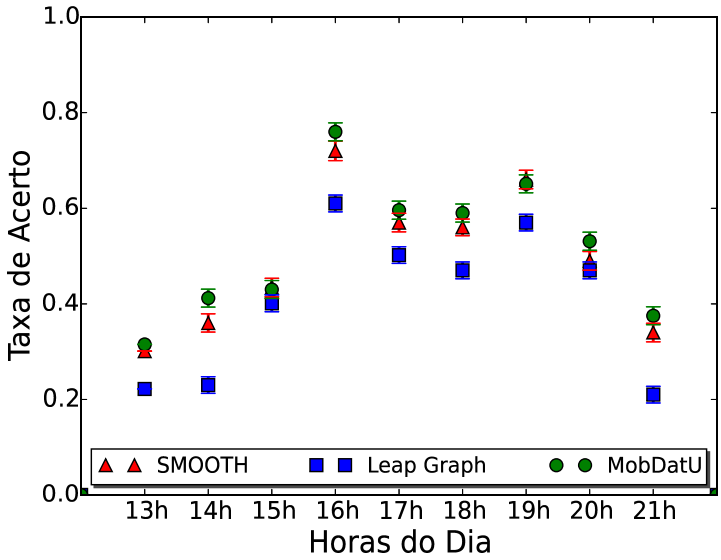


(b) Chamadas para *Tweets*

Figura A.24: Taxas de Acerto Médias por Hora para os Cenário de Dados Heterogêneos - Belo Horizonte Treino e Teste no Dia 22/06/2013

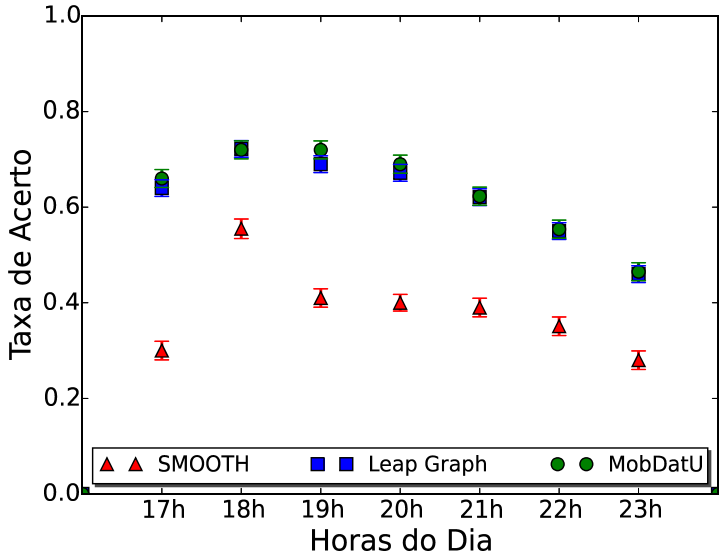


(a) *Tweets* para Chamadas

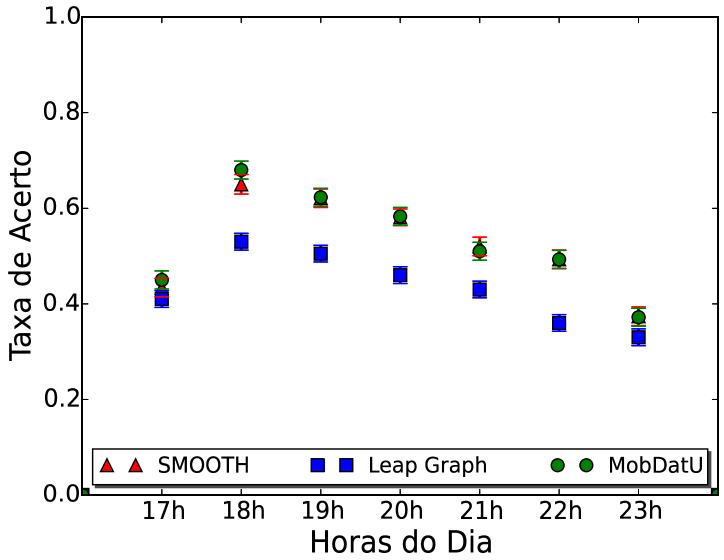


(b) Chamadas para *Tweets*

Figura A.25: Taxas de Acerto Médias por Hora para os Cenário de Dados Heterogêneos - Belo Horizonte Treino e Teste no Dia 26/06/2013

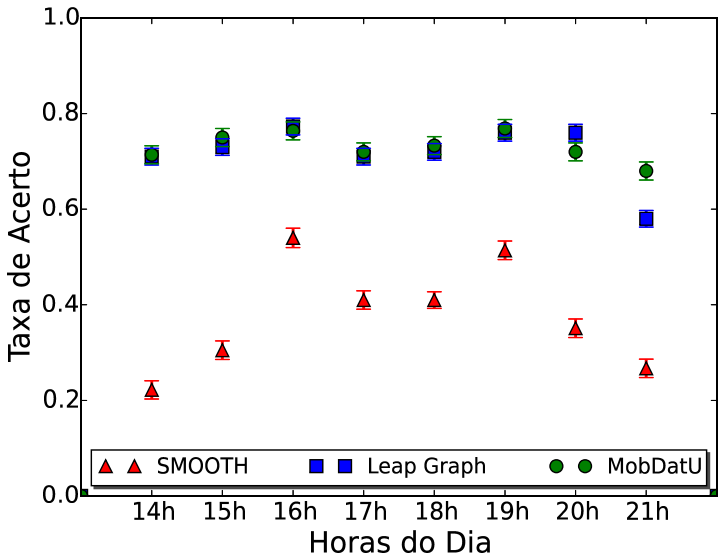


(a) *Tweets* para Chamadas

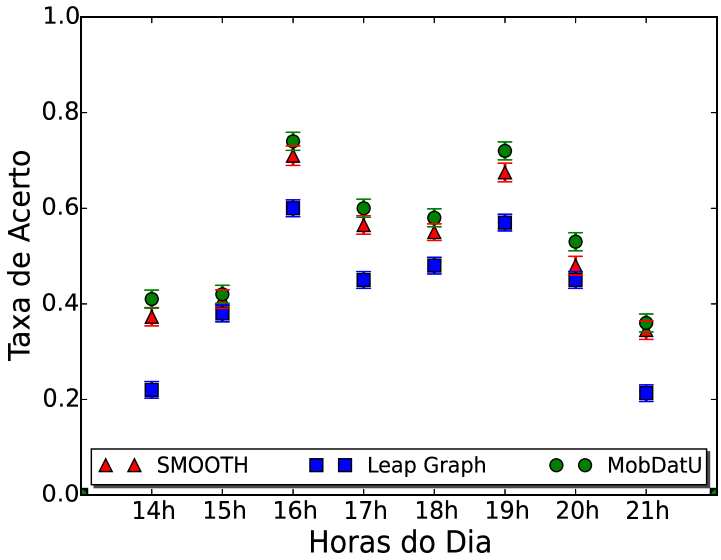


(b) Chamadas para *Tweets*

Figura A.26: Taxas de Acerto Médias por Hora para os Cenário de Dados Heterogêneos - Belo Horizonte Treino e Teste no Dia 11/09/2013

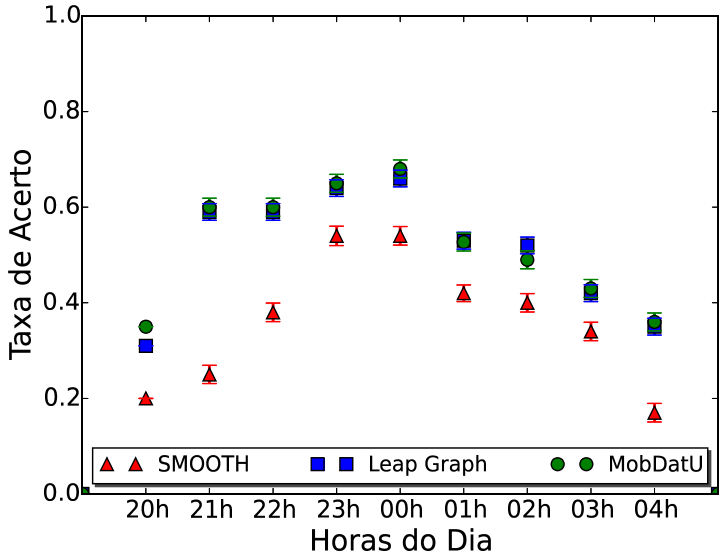


(a) *Tweets* para Chamadas

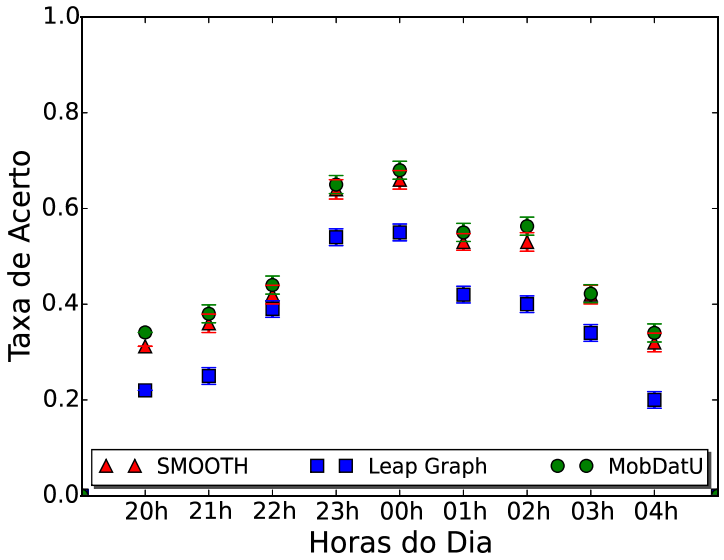


(b) Chamadas para *Tweets*

Figura A.27: Taxas de Acerto Médias por Hora para os Cenário de Dados Heterogêneos - Fortaleza Treino e Teste no Dia 29/06/2014

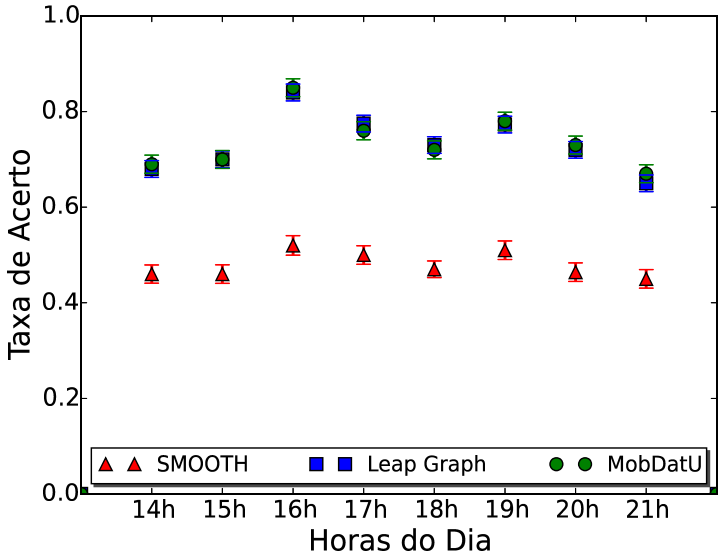


(a) *Tweets* para Chamadas

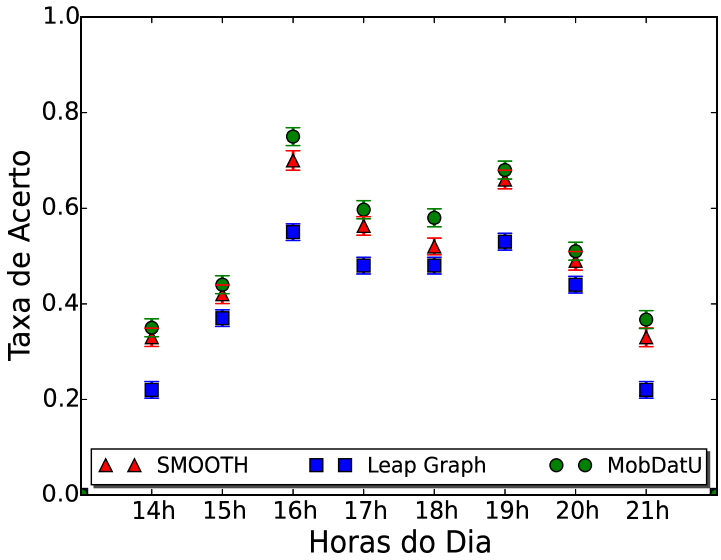


(b) Chamadas para *Tweets*

Figura A.28: Taxas de Acerto Médias por Hora para os Cenário de Dados Heterogêneos - Recife Treino no Dia 31/12/2011 e Teste no Dia 03/01/2012

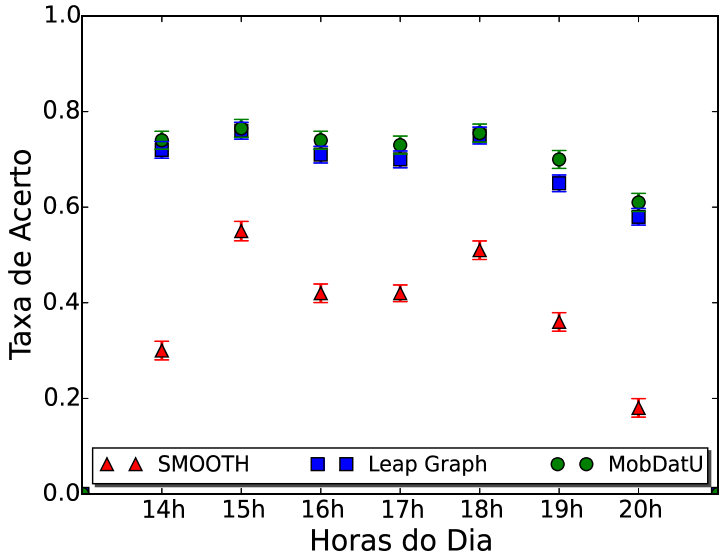


(a) *Tweets* para Chamadas

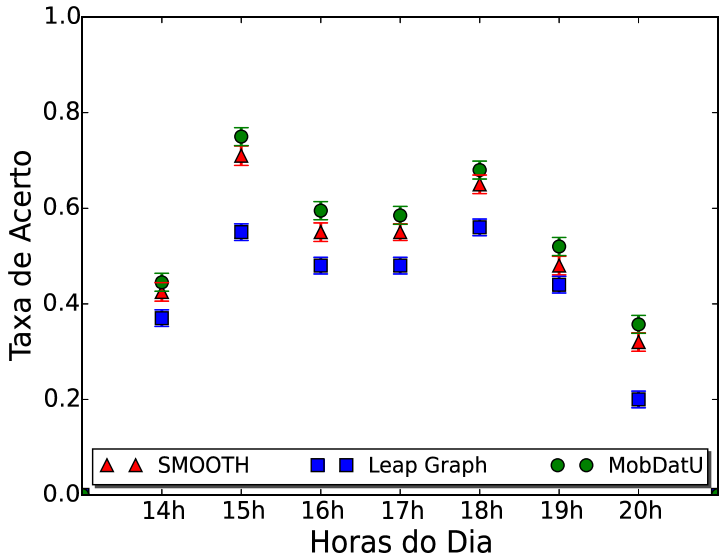


(b) Chamadas para *Tweets*

Figura A.29: Taxas de Acerto Médias por Hora para os Cenário de Dados Heterogêneos - Recife Treino e Teste no Dia 29/06/2014

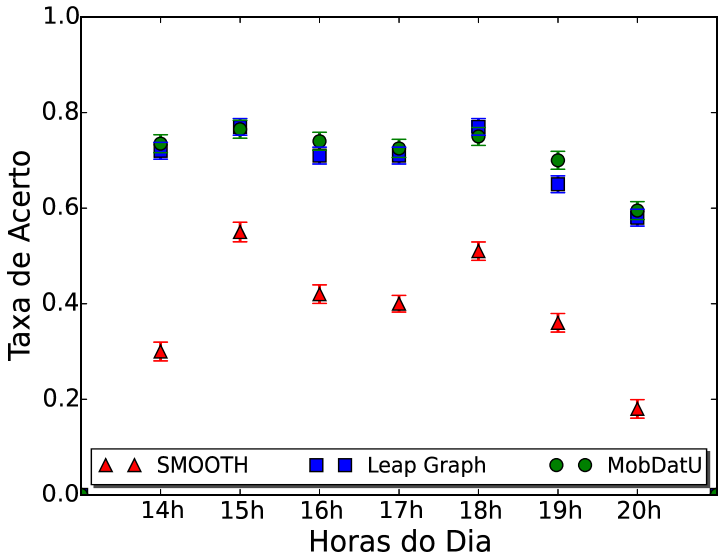


(a) *Tweets* para Chamadas

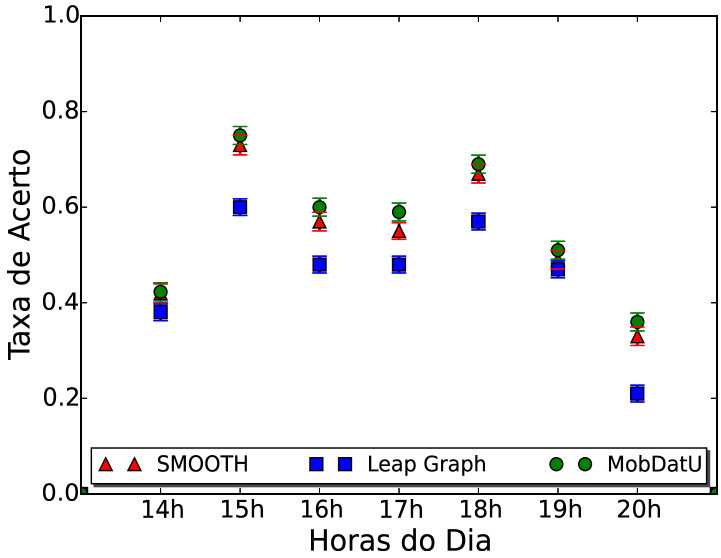


(b) Chamadas para *Tweets*

Figura A.30: Taxas de Acerto Médias por Hora para os Cenário de Dados Heterogêneos - Rio de Janeiro Treino Dia 28/08/2011 e Teste no Dia 30/10/2011

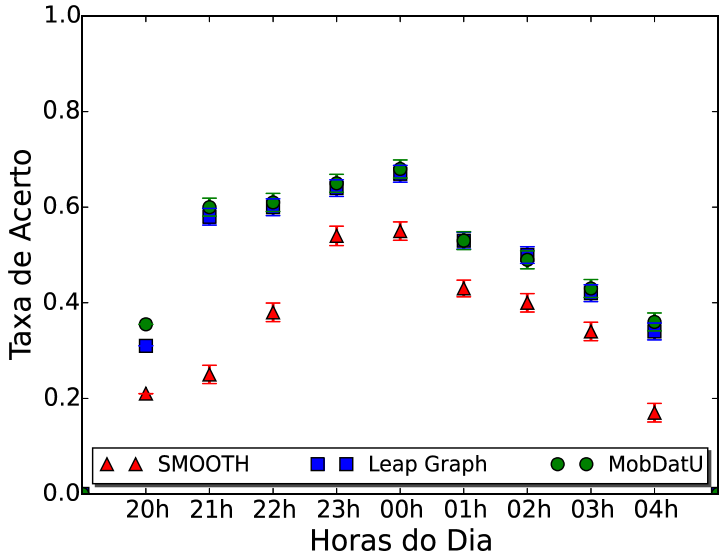


(a) *Tweets* para Chamadas

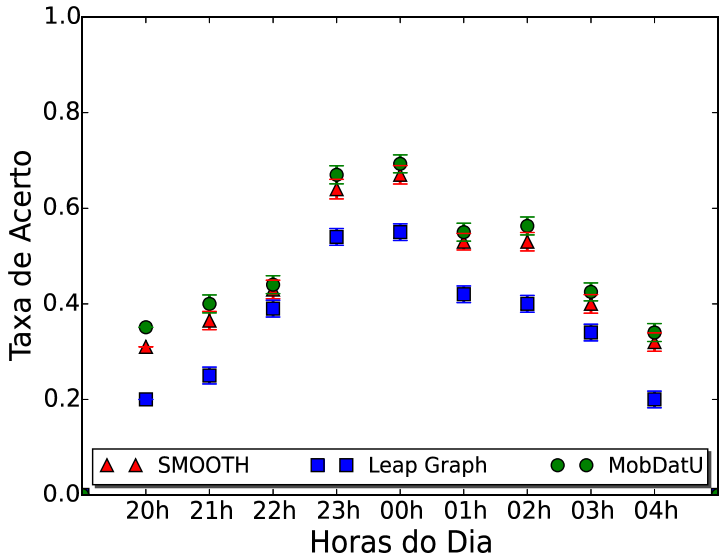


(b) Chamadas para *Tweets*

Figura A.31: Taxas de Acerto Médias por Hora para os Cenário de Dados Heterogêneos - Rio de Janeiro Treino Dia 04/12/2011 e Teste no Dia 11/12/2011

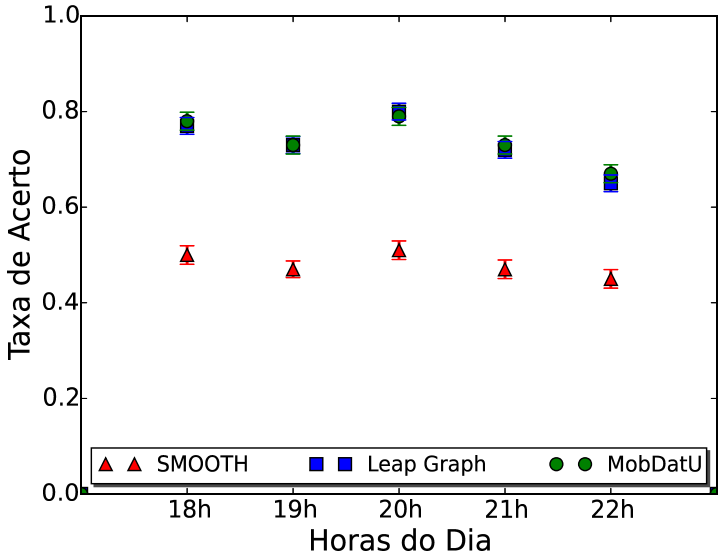


(a) *Tweets* para Chamadas

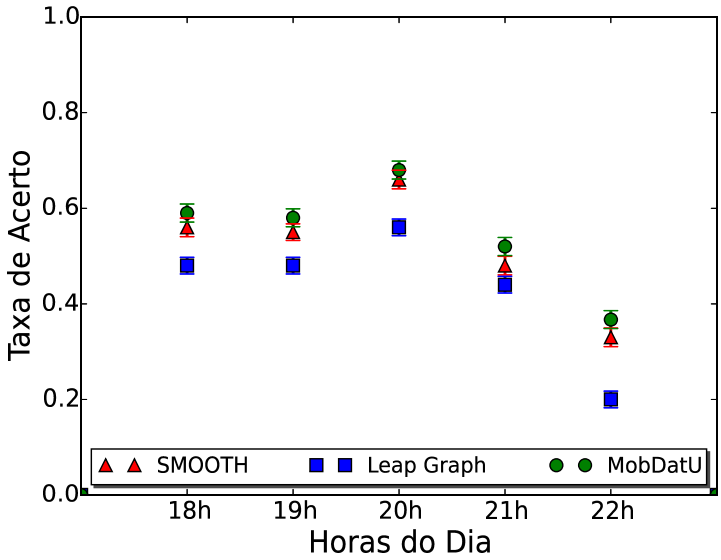


(b) Chamadas para *Tweets*

Figura A.32: Taxas de Acerto Médias por Hora para os Cenário de Dados Heterogêneos - Rio de Janeiro Treino Dia 31/12/2011 e Teste no Dia 03/01/2012

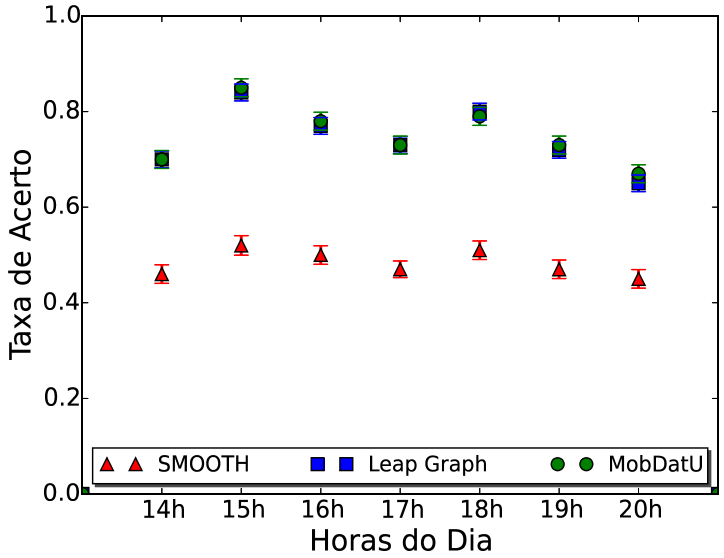


(a) *Tweets* para Chamadas

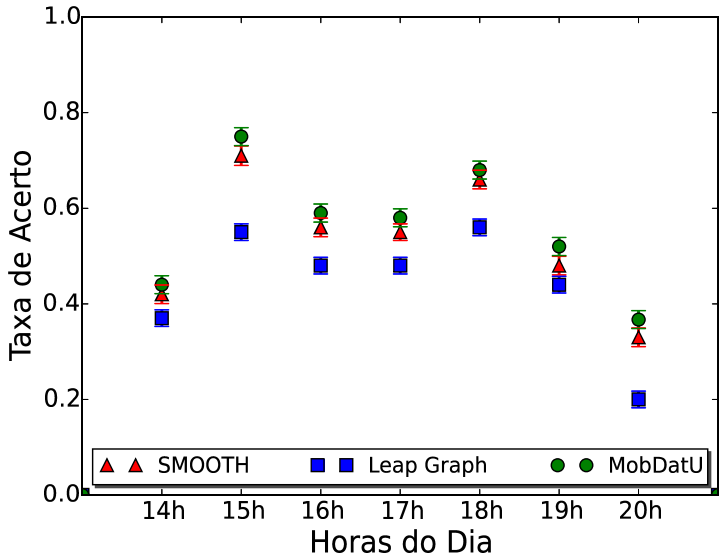


(b) Chamadas para *Tweets*

Figura A.33: Taxas de Acerto Médias por Hora para os Cenário de Dados Heterogêneos - Rio de Janeiro Treino e Teste no Dia 29/03/2012

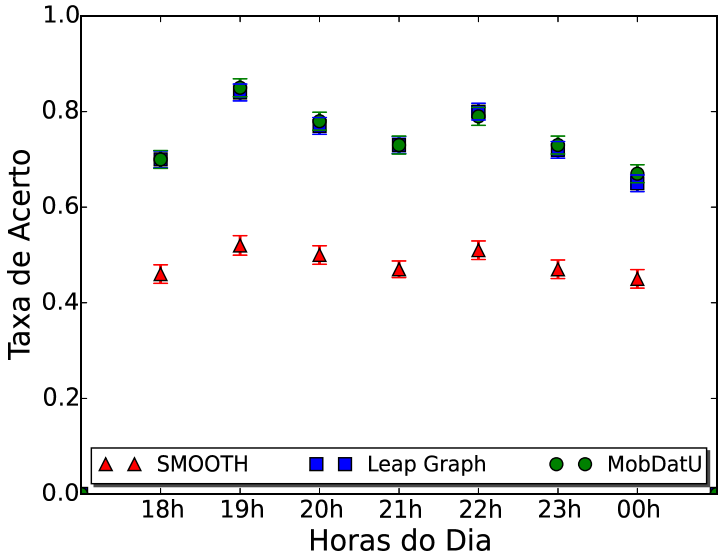


(a) *Tweets* para Chamadas

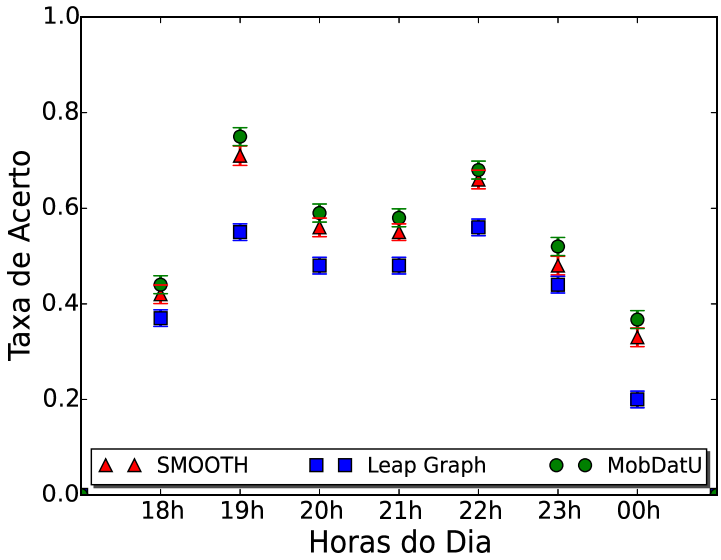


(b) Chamadas para *Tweets*

Figura A.34: Taxas de Acerto Médias por Hora para os Cenário de Dados Heterogêneos - Rio de Janeiro Treino e Teste no Dia 08/07/2012

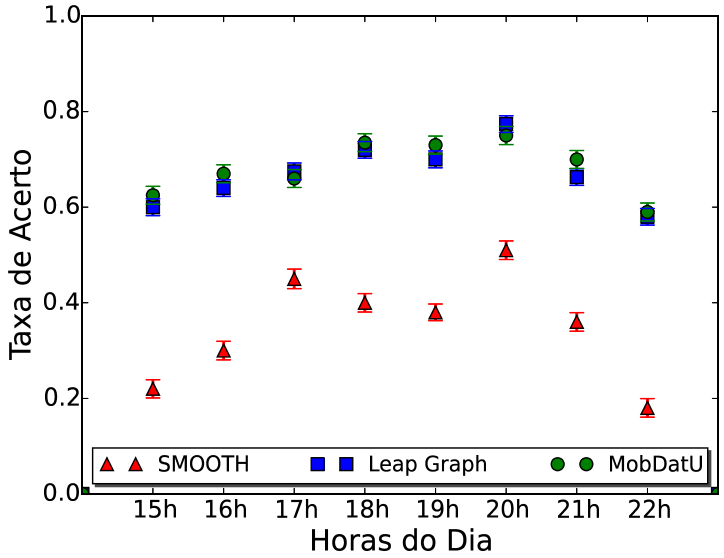


(a) *Tweets* para Chamadas

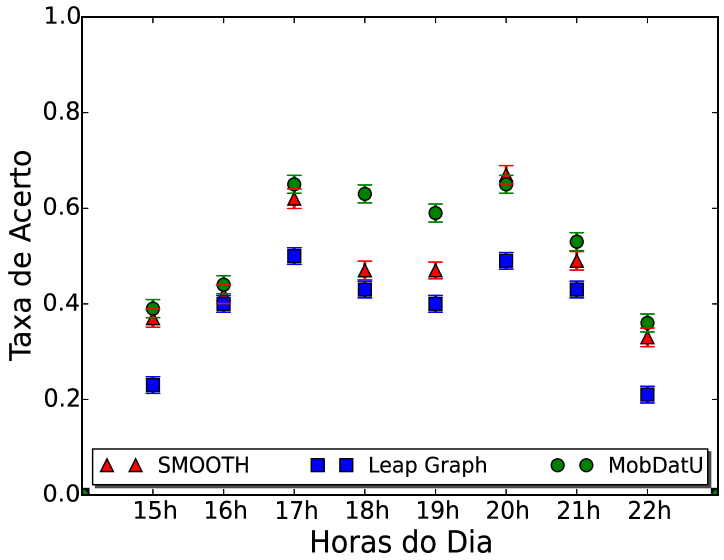


(b) Chamadas para *Tweets*

Figura A.35: Taxas de Acerto Médias por Hora para os Cenário de Dados Heterogêneos - Rio de Janeiro Treino e Teste no Dia 27/11/2013

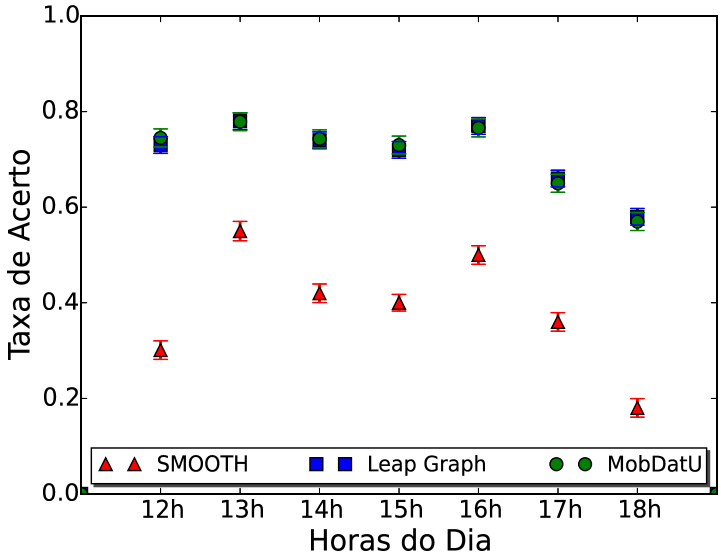


(a) *Tweets* para Chamadas

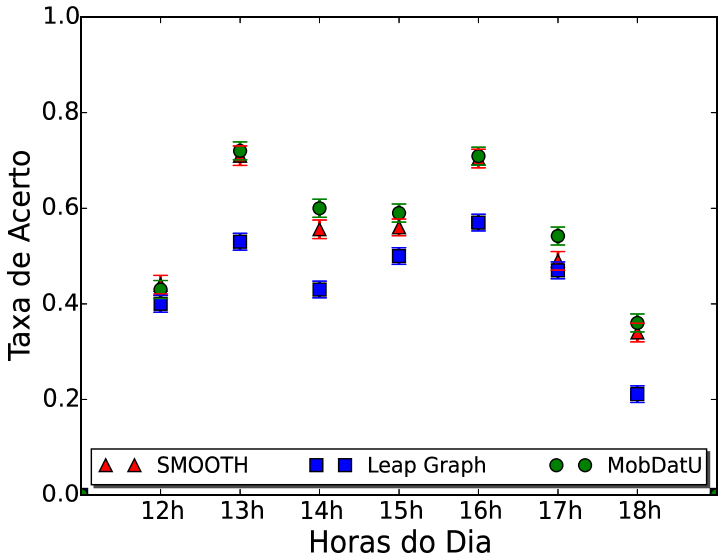


(b) Chamadas para *Tweets*

Figura A.36: Taxas de Acerto Médias por Hora para os Cenário de Dados Heterogêneos - São Paulo Treino e Teste no Dia 04/02/2012

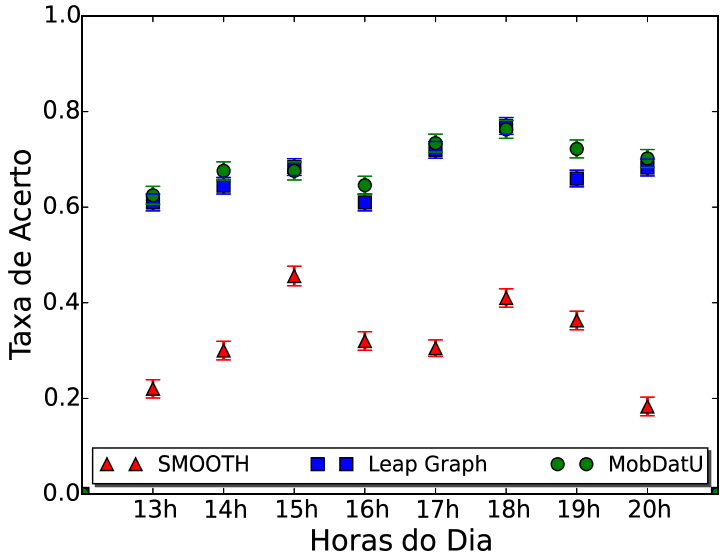


(a) *Tweets* para Chamadas

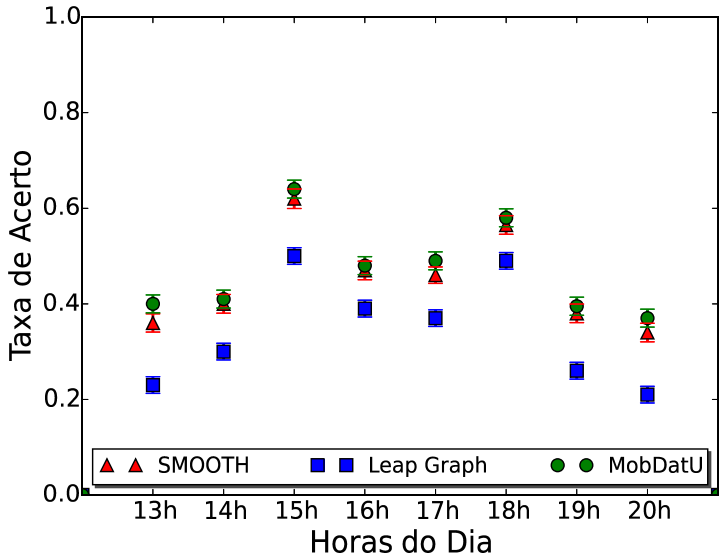


(b) Chamadas para *Tweets*

Figura A.37: Taxas de Acerto Médias por Hora para os Cenário de Dados Heterogêneos - São Paulo Treino e Teste no Dia 25/11/2012



(a) *Tweets* para Chamadas



(b) Chamadas para *Tweets*

Figura A.38: Taxas de Acerto Médias por Hora para os Cenário de Dados Heterogêneos - São Paulo Treino e Teste no Dia 24/03/2013