Pré-processamento e Análise de Dados de Táxis

Cristiano Martins Monteiro[1], **Fábio Rocha da Silva**[1], **Cristina Duarte Murta**[1]

[1]Departamento de Computação – CEFET-MG

Resumo. *O estudo de grandes quantidades de dados é um desafio atual e devemos estar preparados para tratá-las e analisá-las. Nesta tarefa, o pré-processamento é essencial para verificar os dados, identificar inconsistências, possíveis erros e incompletude. Neste trabalho, foram analisadas duas bases de dados com mais de trinta milhões de registros da movimentação de táxis nas cidades de San Francisco e Roma. Propomos um algoritmo para o tratamento das velocidades anômalas identificadas na etapa de pré-processamento destas bases. Apresentamos a análise das bases de dados antes e após a aplicação do algoritmo, mostrando sua relevância e pertinência. Os resultados evidenciam características específicas do serviço de táxi nas duas metrópoles.*

Abstract. *The study of large amounts of data is a current challenge and we must be prepared to treat and analyze them. In this task, pre-processing is essential for verifying data, identifying inconsistencies, possible errors and incompleteness. In this work, two datasets with more than thirty million records of the movement of taxis in the cities of San Francisco and Rome were analyzed. We propose an algorithm to treat anomalous speeds identified in the preprocessing step of these datasets. We present the analysis of the datasets before and after the application of the algorithm, showing its relevance and pertinence. The results show specific characteristics of the taxi service in the two metropolises.*

1. Introdução

A crescente disponibilidade de pegadas digitais de veículos com dispositivos de localização georreferenciada tem possibilitado a análise de padrões da mobilidade urbana, além do estudo de serviços de transporte específicos. Dentre estas pegadas digitais destacam-se as referentes aos meios de transporte público, incentivando pesquisas como integração de dados e descoberta de padrões na mobilidade de ônibus [Kozievitch et al. 2016] e dinâmicas do serviço de táxis [Júnior et al. 2016].

Pegadas digitais dos táxis representam uma importante fonte de dados da mobilidade urbana devido às rotas de táxi não estarem restritas a itinerários e horários fixos, tais como ocorre para linhas de ônibus e metrô. Dada a liberdade de trajetos dos táxis, a identificação do percurso realizado depende das localizações registradas por dispositivos GPS. No entanto, este tipo de dado está sujeito a erros variados, tais como erros de GPS [Valero et al. 2014], processamentos em mapas com vias de trânsito incompletas, incorretas, ou com sentido de circulação equivocado, além de possíveis falhas na aquisição ou armazenamento dos dados. Portanto, os estudos das pegadas digitais dos táxis requerem um pré-processamento dos dados antes de sua análise [Monteiro et al. 2016].

O objetivo deste trabalho é estudar a atividade dos táxis das cidades de San Francisco e Roma, traçando um perfil do uso dos serviços de táxi nestas cidades, a partir de

dados coletados. Para alcançar esse objetivo, foi feito inicialmente o pré-processamento dos dados para identificar possíveis erros e inconsistências. Nesta etapa, foram identificadas distâncias anômalas, e propomos o algoritmo Tratamento de Velocidades Anômalas (TVA), que identifica e ajusta anomalias nas velocidades dos táxis em movimento. Após o pré-processamento, foram analisadas as distâncias percorridas pelos táxis no decorrer de um dia típico em ambas as metrópoles, bem como o uso do serviço ao longo do dia.

Foram tratados e analisados mais de 30 milhões de registros de espaço e de tempo, adquiridos em um mês de circulação dos táxis de ambas as cidades. Consideramos que se trata de um *Big Data* pois o grande volume das bases de dados e o caráter multidimensional de dados espaço-temporais dificultam as análises [Monteiro et al. 2016] e inviabilizam a utilização de algumas técnicas de mineração de dados. Além disso, a própria incerteza da localização dos táxis evidencia o desafio em analisar tal conjunto de dados. Os resultados indicam a viabilidade da metodologia proposta e revelam características específicas dos serviços de táxi em cada cidade estudada. O tratamento proposto pode ser útil para estudos de transportes em geral que analisem as trajetórias de veículos, sem a necessidade de manter mapas precisos das vias da cidade. As análises das distâncias diárias acumuladas dos táxis de San Francisco e Roma podem beneficiar as próprias empresas de táxi, os serviços concorrentes tais como *Uber*, bem como clientes que dependem desse serviço.

Este artigo está organizado em seis seções. A seção seguinte discute os trabalhos relacionados. A Seção 3 apresenta as bases de dados utilizadas neste trabalho e a metodologia. A Seção 4 descreve os algoritmos propostos para o pré-processamento dos dados e analisa seus resultados. A Seção 5 apresenta as análises dos dados após o pré-processamento, e a Seção 6 finaliza o trabalho.

2. Trabalhos Relacionados

Trabalhos recentes utilizam dados de transporte público para analisar padrões do fluxo de veículos e estudar comportamentos dos centros urbanos. Grande parte desses trabalhos se baseiam nas localizações registradas durante o percurso a fim de compreender dinâmicas dos serviços de transporte ou propor melhorias para a mobilidade na cidade.

O uso de técnicas de inferência é comum no estudo de dados obtidos por táxis. Por exemplo, [Ganti et al. 2013] inferem o início e fim das rotas de táxi a partir de padrões identificados nas suas movimentações. Os autores alcançaram precisão superior a 90% utilizando uma medida denominada *Stretch Factor*. Essa medida visa diferenciar os momentos em que o táxi está dando voltas pela cidade (possivelmente sem passageiro) dos momentos em que o táxi está se locomovendo diretamente em um sentido (possivelmente com passageiro). Uma das formas de calcular o *Stretch Factor* se baseia nas distâncias percorridas pelos táxis considerando as vias permitidas das cidades.

Em [Oliveira et al. 2015], os autores avaliam técnicas para a escolha do taxista mais próximo a atender um passageiro, comparando um algoritmo guloso contra um algoritmo de otimização. Um método de cálculo da distância entre o taxista e o passageiro também foi avaliado. Os algoritmos foram comparados utilizando a distância Euclidiana e a distância percorrida pelo táxi considerando o sentido permitido das vias. A conjunção do método de distância com o algoritmo de otimização produziu bons resultados.

Um estudo temporal e espacial do serviço de táxis em Belo Horizonte também foi encontrado na literatura [Júnior et al. 2016]. Neste trabalho, os autores analisaram uma

semana de chamadas, finalizações e cancelamentos de rotas obtidas pelo aplicativo *WayTaxi*. Dentre as análises feitas, foi constatado que 52% das rotas de táxi tiveram distância igual ou menor que dois quilômetros, indicando que a maior parte dos usuários do aplicativo na cidade não solicita trajetos longos de táxi.

Tratamentos das localizações e distâncias percorridas pelos veículos são importantes durante o pré-processamento dos dados. Tais tratamentos são úteis principalmente ao estudar a mobilidade de regiões com vias próximas às outras, conforme demonstrado em [Jones et al. 2007]. Possíveis erros de GPS ao localizar o percurso do veículo podem resultar no mapeamento irreal das vias percorridas. Para contornar esse problema, [Jones et al. 2007] propõem algoritmos para identificar as vias utilizadas por um veículo dado as localizações registradas e a rede de estradas da região. Porém, esses algoritmos não se aplicam às distâncias calculadas entre trechos, tais como as obtidas por meio das ferramentas *Google Maps Distance Matrix API*, *Bing Maps Rest Services – Routes API*, e *Here REST APIs – Calculate Matrix*, dentre outras.

As distâncias das rotas de táxi também foram estudadas em [Monteiro et al. 2016]. Além de padrões espaciais e temporais da mobilidade de táxis em duas metrópoles, foram analisadas as distâncias das rotas de táxi em San Francisco após o tratamento de táxis parados, proposto inicialmente naquele artigo. O presente trabalho estende este trabalho anterior dos mesmos autores, e se diferencia dos trabalhos relacionados por propor o Tratamento de Velocidades Anômalas, e por analisar o funcionamento dos táxis de San Francisco e de Roma com base nas distâncias diárias acumuladas. Estas distâncias diárias foram obtidas por ferramentas de distâncias calculadas entre trechos, e suas velocidades extremas foram corrigidas aplicando o tratamento proposto neste artigo.

3. Bases de Dados e Metodologia

Neste artigo utilizamos duas bases de dados que registram deslocamentos de táxis nas cidades de San Francisco, EUA, e Roma, Itália. Não encontramos bases de dados similares coletadas em metrópoles brasileiras. A base de dados de San Francisco[1] é composta por 536 arquivos texto, cada arquivo referente a um táxi [Piorkowski et al. 2009]. No total, a base contém 11.219.955 linhas, e cada linha registra os seguintes dados: identificação do táxi; localização em latitude e longitude; *status* de ocupação (1 para táxi com passageiro e 0 para táxi sem passageiro); e *timestamp* no formato *Unix Epoch* do momento de aquisição destes dados. A base de dados de Roma[2], contém 21.817.851 registros em um único arquivo sobre 316 taxistas diferentes [Bracciale et al. 2014]. Cada linha contém a identificação do taxista, a localização em latitude e longitude, e a data e a hora de aquisição dos dados. Os dados da base de San Francisco foram armazenados a cada 60 segundos em média, de 17/05/2008 a 10/06/2008, enquanto os dados da base de Roma foram armazenados a cada 7 segundos em média, de 01/02/2014 a 03/03/2014.

É importante ressaltar que a base de dados de San Francisco teve suas informações registradas por meio de um dispositivo acoplado aos táxis, enquanto a base de Roma é obtida por meio de *tablets* que estavam com os taxistas. Além disso, a base de dados de Roma não informa quando o taxista estava com ou sem passageiro. Utilizamos o termo "registro" para referir a cada linha das bases de dados; "trecho" para definir o movimento

[1] http://crawdad.org/epfl/mobility/20090224/

[2] http://crawdad.org/roma/taxi/20140717/

do táxi a cada dois registros consecutivos; e "rota" para representar uma sequência de registros com o mesmo *status* de ocupação. Como apenas a base de San Francisco registra a ocupação dos táxis, discutimos rotas somente para San Francisco.

As distâncias percorridas pelos táxis foram calculadas utilizando a *Google Maps Distance Matrix API*. Essa *API* foi escolhida após comparação com várias APIs disponíveis, ver detalhamento das opções em [Monteiro 2016]. Esta *API* considera o sentido permitido das vias de trânsito e permite até 2.500 consultas gratuitas, diariamente, de distâncias entre trechos, para cada conta aberta no sistema. No nosso caso, foram utilizadas cerca de duas centenas de contas para que o resultado apresentado aqui fosse obtido. Devido à grande quantidade de trechos nas duas bases de dados, obter a distância percorrida para todos os trechos sem algum pré-processamento seria inviável dado o limite de consultas gratuitas. Este problema foi solucionado conforme descrito a seguir.

Optamos por reduzir a precisão das coordenadas geográficas registradas visando diminuir a quantidade de localizações únicas, mesmo procedimento adotado em [Rossi et al. 2015, Monteiro et al. 2016]. A precisão das coordenadas com quatro casas decimais é de 11,132 metros na linha do Equador, consideramos essa precisão aceitável para localizar um automóvel. Portanto, as coordenadas geográficas de ambas as bases de dados foram arredondadas para quatro casas decimais e não foram feitas consultas repetidas à *API* do *Google Maps* para trechos que, após o arredondamento, tenham o mesmo local de início e fim. Dessa forma, a quantidade de trechos a consultar para a base de San Francisco foi reduzida de 11.218.651 para 7.351.320 (queda de 34,47%), e para a base de Roma foi reduzida de 21.817.828 para 4.515.642 (queda de 79,30%). A redução maior em Roma é explicada pelo menor tempo médio entre coletas de dados (sete segundos).

Ambas as bases de dados foram filtradas para retirar registros que apresentaram evidências de erros de localização, incluindo registros de táxis localizados no mar, trechos de centenas de quilômetros que teriam sido percorridos em segundos, ou de trechos em que a *API* do *Google Maps* não encontrou uma rota permitida. Estes filtros retiraram menos de 0,008% dos dados de San Francisco e Roma. A retirada destes registros anômalos foi discutida em [Monteiro et al. 2016]. Após estes pré-processamentos, as distâncias entre os trechos foram coletadas utilizando-se a API escolhida. A partir destes dados, aplicamos os algoritmos Tratamento de Táxis Parados [Monteiro et al. 2016] e Tratamento de Velocidades Anômalas (apresentado na subseção 4.2). Finalmente, foram realizadas análises estatísticas e exploratórias das distâncias, apresentadas nas seções 5.1 e 5.2.

Após estes processamentos, avaliamos se há diferença com 5% de significância entre as distâncias calculadas antes e após o Tratamento de Velocidades Anômalas, utilizando o teste estatístico de Kolmogorov-Smirnov para duas amostras. Este teste foi escolhido por ser não paramétrico e livre de distribuição, sendo assim mais robusto [Gibbons and Chakraborti 2003].

4. Algoritmos para o Tratamento dos Dados

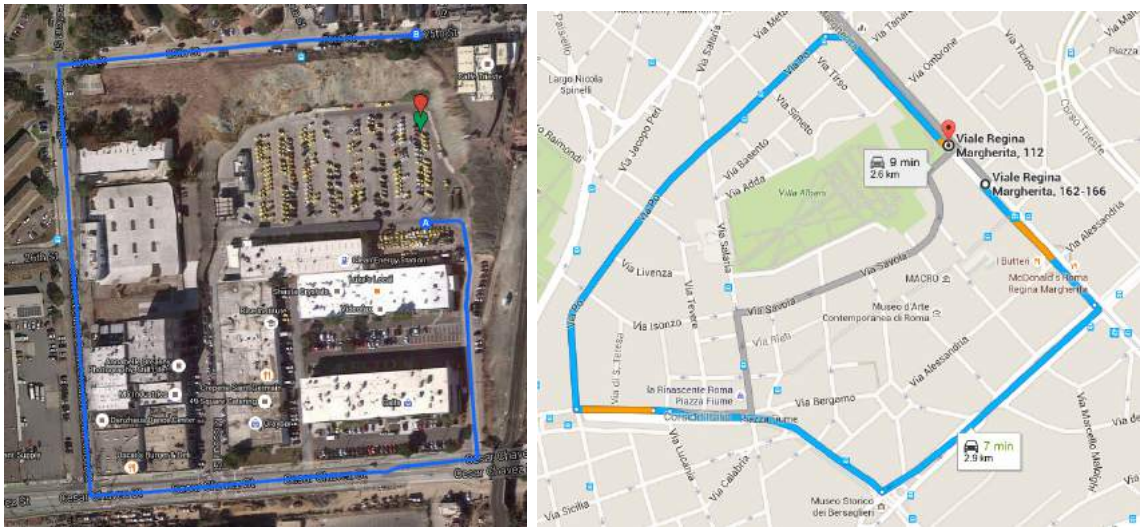
Esta seção aborda os algoritmos utilizados para o tratamento das distâncias percorridas pelos táxis em San Francisco e Roma. Inicialmente discutimos as distâncias anômalas obtidas pela *API* do *Google Maps* e o tratamento já existente para eliminar parte destas anomalias. Finalmente, apresentamos o tratamento proposto para corrigir as velocidades

extremas identificadas.

4.1. Distâncias Anômalas Identificadas

As distâncias percorridas pelos táxis são importantes, por exemplo, para análises exploratórias das dinâmicas da cidade [Alvarenga et al. 2016, Júnior et al. 2016], e para a mineração de trajetórias [Ganti et al. 2013]. Identificar e tratar anomalias no cálculo dessas distâncias é crucial para a correção de possíveis erros nas bases de dados [Monteiro et al. 2016].

As distâncias de todos os trechos percorridos pelos táxis foram calculadas utilizando-se a *Google Maps Distance Matrix API*. Essa ferramenta recebe como parâmetro os locais de início e fim de um trecho (além de outros parâmetros opcionais) e retorna a distância e duração estimadas para o percurso utilizando as vias permitidas. Porém, há situações em que a distância estimada para um trecho é muito acima do esperado. A título de exemplo, mostramos duas situações identificadas como anomalias, calculadas pela citada *API*, e ilustradas na Figura 1.



(a) Distância com o táxi parado (b) Distância com o táxi movimentando

Figura 1. Distâncias anômalas calculadas pelo *Google Maps*

A Figura 1 (a) mostra o estacionamento da empresa *SF Yellow Cab*, na cidade de San Francisco, e ilustra uma distância anômala calculada neste estacionamento. Alguns táxis permaneciam, aparentemente, parados neste estacionamento por horas, sem interromper o armazenamento de registros da sua localização e tempo. Em diversos momentos, a localização do táxi variava poucos metros em torno de um mesmo ponto. Esse comportamento é ilustrado na Figura 1 (a) pelo deslocamento de um táxi do local marcado em verde ao local marcado em vermelho. Essas pequenas variações registradas em torno de um mesmo local causam anomalias significativas nas análises de mobilidade dos táxis, principalmente quando a distância calculada considera o sentido permitido das vias de trânsito. A distância do trecho entre o ponto verde e o ponto vermelho indicados na figura é de somente 12 metros. Porém, o *Google Maps* estimou que o táxi teria percorrido um trecho de 958 metros (ilustrado em azul) do ponto verde ao ponto vermelho. Anomalias deste tipo foram tratadas pelo algoritmo Tratamento de Táxis Parados (TTP), proposto em [Monteiro et al. 2016].

A Figura 1 (b) ilustra uma distância anômala quando um táxi de Roma estava movimentando na avenida *Viale Regina Margherita*. A localização do táxi variou somente 109 metros (do ponto branco até a marcação em vermelho). Porém, a *API* do *Google Maps* estimou que o táxi teria percorrido uma distância de 2,9 quilômetros de extensão (em azul). Uma vez que o intervalo de tempo deste trecho foi de apenas 1,2 segundo, a velocidade do táxi ao percorrer os 2,9 quilômetros estimados teria sido de 8.619 km/h. É possível que, devido a um erro de GPS, o táxi tenha sido localizado na contra-mão, fazendo o *Google Maps* propor um retorno muito maior seguindo as vias de trânsito permitidas. Também é possível que a localização do táxi esteja correta, mas que haja um erro nas vias catalogadas no mapa ou uma imprecisão da *API* que calcula as distâncias. Este tipo de anomalia foi tratado pelo algoritmo Tratamento de Velocidades Anômalas, apresentado na subseção a seguir.

4.2. Tratamento de Velocidades Anômalas

Esta subseção apresenta o tratamento proposto para situações em que foram identificadas velocidades anômalas para um táxi em movimento, como por exemplo, ilustrado na Figura 1 (b). O Tratamento de Velocidades Anômalas (TVA) pode ser aplicado em análises que envolvam cálculo de distâncias e velocidades de veículos, bem como em análises de objetos móveis que estejam propensos a erros de localização, de registro de tempo ou erros do próprio mapa.

Para calcular a distância percorrida em um trecho, a *API* do *Google Maps* recebe como parâmetros de localização somente os pontos de início e fim do trecho. Esta *API* utiliza um algoritmo similar ao "Simple Distance Map Matching (SDMM)", apresentado em [Jones et al. 2007], para ajustar cada localização de GPS a uma via, dado que a identificação destes pontos no mapa aparenta ser baseada somente na rua mais próxima.

O TVA pode ser descrito da seguinte forma. Considere que um táxi está seguindo uma rota, composta por uma sequência de coordenadas: $c_1, c_2, c_3, c_4, ...$ A *API* recebe inicialmente o trecho (c_1, c_2), e retorna sua distância, que é dividida pela diferença entre os tempos registrados em cada coordenada, produzindo assim a velocidade estimada para o trecho. Uma velocidade é considerada anômala sempre que ultrapassar um limiar v. Toda vez que uma velocidade anômala é calculada em um trecho, o algoritmo TVA refaz o cálculo da distância tomando como destino a coordenada seguinte. Por exemplo, se for calculada uma velocidade anômala para o trecho (c_1, c_2), o algoritmo tomará como destino a coordenada c_3, e assim calculará nova velocidade para o trecho (c_1, c_3). Se o problema persistir, o algoritmo tomará como destino a coordenada c_4, e calculará nova velocidade para o trecho (c_1, c_4), e assim por diante, até que a velocidade obtida seja menor ou igual que v ou não haja mais coordenadas aceitáveis na sequência. A premissa é que erros pontuais como imprecisões do registro de tempo ou do GPS ou mesmo da API poderão ser eliminados facilmente.

A técnica foi aplicada aos dados após o tratamento do algoritmo TTP, utilizando o limiar v = 150 km/h. Depois de aplicar o TVA, o número de trechos na base de dados de San Francisco foi reduzido em 2,82%. Na base de dados de Roma a redução foi de 10,58%. A menor identificação de trechos anômalos em San Francisco se deve ao fato de que o tempo entre coletas de dados nesta base é de um minuto. Portanto, somente trechos com distância acima de 2,5 quilômetros (supostamente percorridos em um minuto) indicarão velocidade maior que o limiar v = 150 km/h. Já em Roma, os registros foram feitos

em intervalos mais curtos, de sete segundos em média, o que amplifica potenciais erros, e explica o maior impacto da aplicação do TVA.

Uma vez que não é possível saber quando houve erro no cálculo da distância, não é possível calcular medidas como *precision*, *recall* e acurácia do TVA. Assim, avaliamos se houve diferença significativa entre as distâncias antes e após a aplicação do TVA. O teste estatístico de Kolmogorov-Smirnov foi utilizado para avaliar a hipótese nula de igualdade entre as seguintes distâncias antes e após a aplicação do TVA: (i) todas as distâncias dos trechos; (ii) todas as distâncias das rotas; (iii) somente entre rotas de táxis com passageiro; (iv) somente entre rotas de táxis sem passageiro; e (v) entre as distâncias diárias acumuladas dos táxis. Este teste estatístico é importante para certificar que o tratamento promoveu diferença relevante nas distâncias e velocidades.

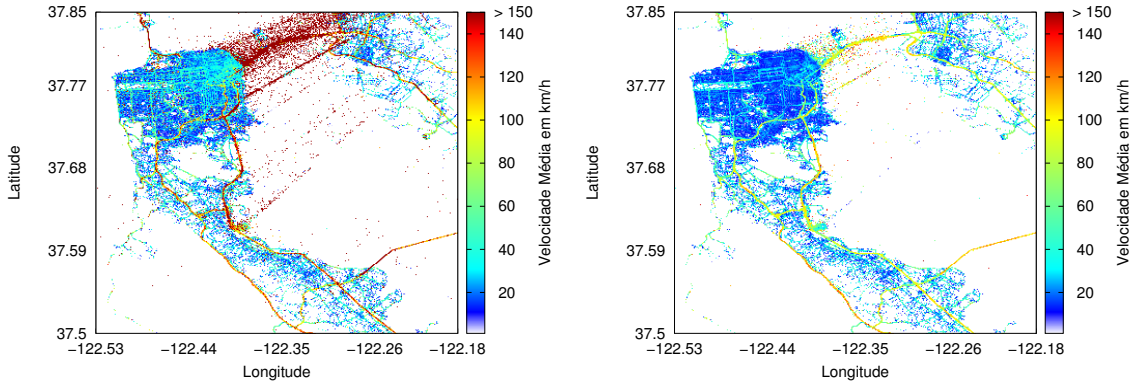
Para todos estes casos, na base de dados de San Francisco, o p-valor foi menor que $2{,}2\times10^{-16}$ (menor valor de arredondamento do *software R*). Portanto, podemos afirmar com 5% de significância que há evidências estatísticas para refutar a hipótese de igualdade entre as distribuições. Para a base de dados de Roma, obtivemos o mesmo resultado para os casos (i) e (v). Os demais casos não foram avaliados para esta base porque não há a informação de status de ocupação para a definição das rotas. Os resultados evidenciam que o algoritmo TVA tem um impacto significativo no cálculo de distâncias de cada trecho, das distâncias das rotas percorridas com ou sem passageiro em San Francisco e das distâncias acumuladas no decorrer do dia. Comparações estatísticas adicionais a respeito do impacto dos algoritmos TTP e TVA são apresentadas em [Monteiro 2016]. Tais análises foram omitidas neste trabalho devido à restrição de espaço.

Os resultados da aplicação do TVA nas bases de dados de San Francisco e de Roma são apresentados a seguir. A Figura 2 apresenta mapas de calor após a aplicação somente do TTP e após a aplicação do TTP e TVA na cidade de San Francisco. Um mapa de calor é uma representação gráfica de pontos em uma matriz colorida. Cada ponto deste mapa representa a velocidade média dos táxis em determinado local da cidade. Quanto mais vermelho o ponto, maior é a velocidade média registrada no local. A Figura 2 (a) apresenta as velocidades médias antes da aplicação do TVA e a Figura 2 (b) apresenta as velocidades médias após a aplicação do TVA. Os locais em que a velocidade média ultrapassava 150 km/h antes do TVA tiveram suas velocidades médias reduzidas para um valor em torno de 100 km/h após TVA. Esse valor condiz com o limite de velocidade nesta cidade, que é de 70 milhas por hora[3], o equivalente a 112,65 km/h.

A Figura 3 apresenta os mesmos resultados para a base de dados de Roma. Após o TVA, as velocidades médias das vias de trânsito rápido em Roma reduziram de mais de 150 km/h para velocidades em torno de 80 km/h a 100 km/h. Esses valores condizem com o limite de velocidade da cidade, que é de 130 km/h[4]. Na região central, as velocidades médias reduziram também de valores acima de 150 km/h para velocidades em torno de 20 km/h a 40 km/h, ficando similares às da região central de San Francisco e razoáveis para táxis que trafegam no centro de uma cidade. Estes resultados indicam que o algoritmo TVA identifica e reduz anomalias nas distâncias e velocidades em bases de dados de mobilidade urbana. Estas anomalias podem ocorrer devido a diversas fontes de erros e o tratamento é necessário para contornar valores improváveis no contexto estudado.

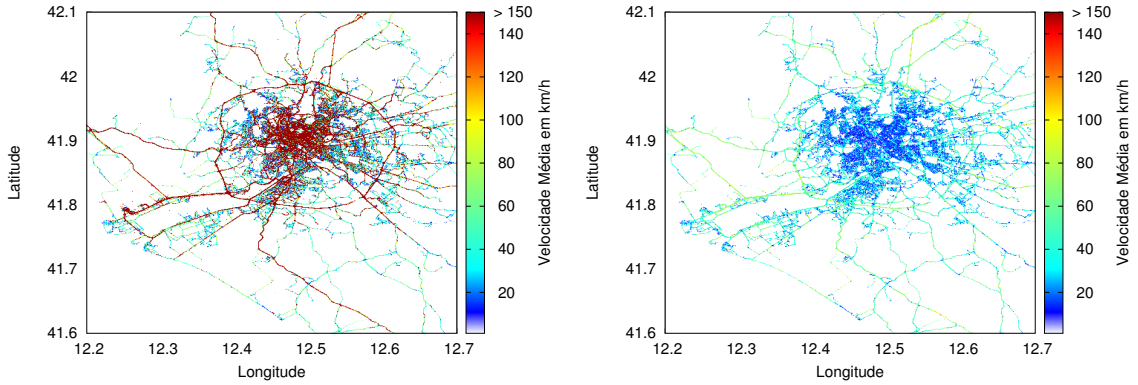
[3] http://carrentalscout.com/driving-speed-limits-san-francisco

[4] http://carrentalscout.com/driving-speed-limits-rome



(a) Tratamento de Táxis Parados (b) Tratamento de Velocidades Anômalas

Figura 2. Impacto do algoritmo TVA em San Francisco



(a) Tratamento de Táxis Parados (b) Tratamento de Velocidades Anômalas

Figura 3. Impacto do algoritmo TVA em Roma

5. Resultados e Comparação entre os Serviços de Táxis das Metrópoles

Esta seção apresenta a análise das distâncias diárias e velocidades calculadas para San Francisco e Roma após a aplicação dos algoritmos TTP e TVA. A subseção 5.1 aborda os resultados para San Francisco. A subseção 5.2 aborda a análise da base de dados de Roma. Esta análise se diferencia das apresentadas em [Alvarenga et al. 2016, Monteiro et al. 2016] por considerar as distâncias percorridas e não a quantidade de táxis ou o número de registros obtidos. Dessa forma, táxis estacionados por horas (ou dias) não interferem nas análises do horário de atividade dos serviços de táxi, por exemplo.

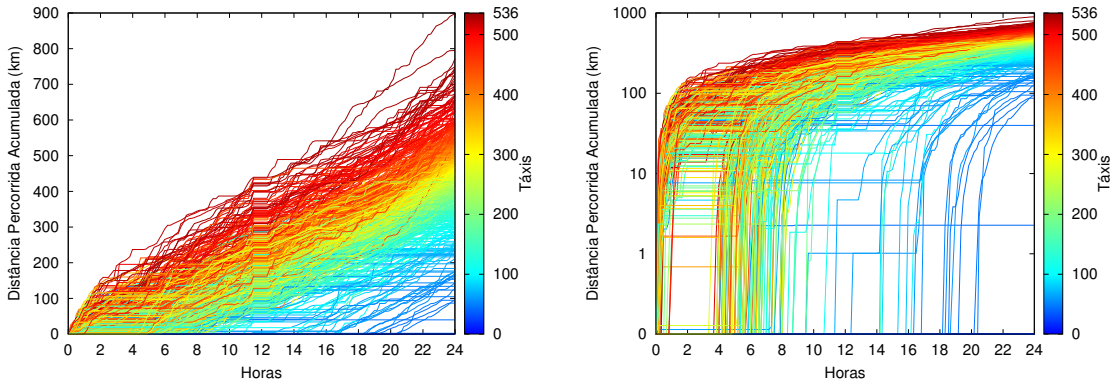
Preocupar-se com tais interferências é necessário dado que, após o TTP, 23,28% dos trechos de San Francisco foram inferidos como parados. Em Roma, 56,54% dos trechos foram inferidos como parados. O fato de mais da metade dos táxis da base de dados de Roma estarem parados condiz com as informações dos guias turísticos de Roma[5][6][7] os quais mencionam que os táxis da cidade normalmente não trafegam pelas ruas procurando passageiros. Nesse caso, é recomendado aos passageiros que se direcionem a um ponto de táxi ou que telefonem para uma companhia de táxis solicitando o serviço.

[5]http://wikitravel.org/en/Rome
[6]http://www.rome.info/transportation/
[7]http://europeforvisitors.com/rome/transportation/rome-taxis.htm

5.1. Análise das Distâncias e Velocidades dos Táxis em San Francisco

Esta seção apresenta a análise das distâncias diárias percorridas pelos táxis de San Francisco, bem como de suas velocidades. Compreender a evolução da distância trafegada pelos táxis no decorrer do dia é importante para o gerenciamento das agências de táxi e para o planejamento de transportes públicos com base nos horários que os táxis estavam em atividade, por exemplo.

A Figura 4 apresenta as distâncias acumuladas pelos táxis de San Francisco durante um dia típico (21/05/2008, quarta-feira, não feriado). Vários dias foram analisados e apresentaram resultados similares [Monteiro 2016]. Cada linha representa um táxi, sendo que as linhas mais vermelhas representam os táxis que percorreram distâncias maiores no decorrer do dia, e as linhas mais azuis representam os táxis que percorreram distâncias menores no decorrer do dia. Na Figura 4 (a) nota-se que, em geral, as distâncias acumuladas ao longo das horas crescem em ritmo linear. O táxi com maior distância percorrida alcançou cerca de 898 quilômetros percorridos até o final do dia. Essa distância é aceitável, considerando que mais de um taxista pode ter dirigido o mesmo táxi no dia e em turnos diferentes.



(a) Distâncias após TVA (b) Distâncias em escala logarítmica

Figura 4. Distâncias acumuladas no dia 21/05/2008 em San Francisco

Observa-se que as distâncias percorridas pela maioria dos táxis aumentam no decorrer do dia, com crescimento pequeno ou nulo no início da madrugada, indicando menor quantidade de táxis circulando neste horário. Em todos os táxis de San Francisco no dia em questão, não foram encontrados registros em torno das 12:00. Esta ausência de registros é representada nos gráficos da Figura 4 pela pausa no aumento das distâncias.

A Figura 4 (b) apresenta o mesmo gráfico da Figura 4 (a), porém, com escala logarítmica no eixo y. Observa-se que um grupo de táxis começou a registrar distâncias entre as 04:00 e 08:00 da manhã. Por outro lado, não houve táxis iniciando as suas movimentações no período das 01:00 às 03:00 da manhã ou a partir das 21:00. Temos como hipótese de que os táxis que teriam iniciado as movimentações às 00:00 são táxis que já estavam em serviço desde o dia anterior. Portanto, parece ser incomum um táxi iniciar suas atividades no período das 21:00 às 03:00 da manhã em San Francisco.

A Tabela 1 apresenta medidas estatísticas das distâncias diárias percorridas pelos táxis de San Francisco, e também das velocidades calculadas após a aplicação do TTP e após a aplicação do TVA. Observa-se que a aplicação do TVA produz uma pequena queda

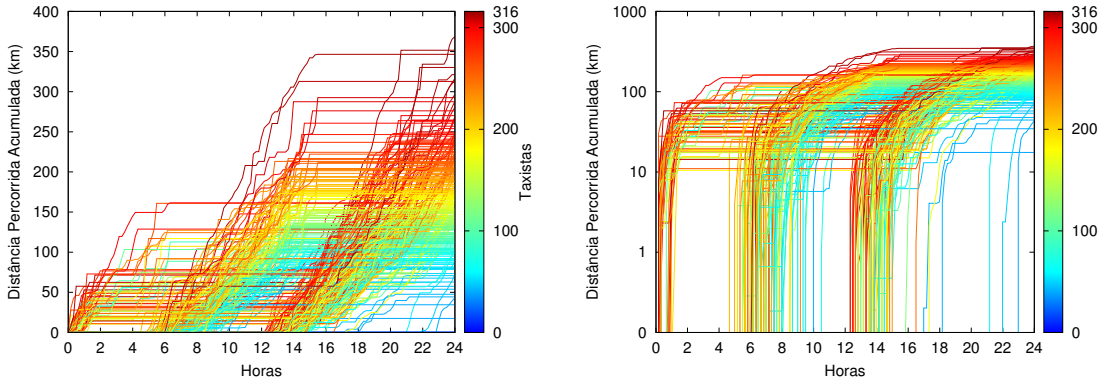
nas medidas apresentadas, exceto nas medidas da cauda (percentis e maior valor). Em especial, o maior valor é bastante reduzido, afetando a média. Os dados consideram todos os dias da base de dados de San Francisco. A maior distância diária obtida foi de 970,79 quilômetros, valor 21% menor que o percentil 99 (de 767,87 quilômetros). A velocidade média é compatível com o funcionamento dos táxis, os quais podem ficar estacionados por horas ou circular pela cidade em velocidades reduzidas procurando por passageiros.

Tabela 1. Distâncias diárias e velocidades em San Francisco

	Distâncias Diárias (km)		Velocidades (km/h)	
Medidas	após TTP	após TVA	após TTP	após TVA
1º Quartil	403,896	359,28	0,68	0,67
Mediana	520,753	454,43	18,44	18,35
Média	543,532	441,46	29,66	26,29
3º Quartil	638,130	538,96	34,38	34,08
Percentil 90	757,389	617,96	73,07	69,79
Percentil 99	1.031,911	767,87	145,49	128,83
Maior	17.155,876	970,79	46.703,83	150,00
Coef. de variação	0,99	0,34	2,48	1,20

5.2. Análise das Distâncias e Velocidades dos Táxi em Roma

A Figura 5 (a) apresenta as distâncias diárias acumuladas em um dia típico na cidade de Roma (12/02/2014, quarta-feira, não feriado). Diferentemente de San Francisco, observa-se aqui mais claramente um padrão de uso do serviço de táxi, em que há basicamente dois grupos de taxistas iniciando seu trabalho ao longo do dia. Um grupo inicia o trabalho na parte da manhã, e outro inicia a partir das 12:00, sendo que alguns taxistas estendem sua jornada de trabalho até após a meia noite.



(a) Distâncias após TVA (b) Distâncias em escala logarítmica

Figura 5. Distâncias acumuladas no dia 12/02/2014 em Roma

A Figura 5 (b) apresenta os mesmos dados que a Figura 5 (a), porém com o eixo y em escala logarítmica para realçar o início das jornadas. Pode-se notar dois grupos de taxistas iniciando a jornada de trabalho: um grupo das 05:00 às 10:00 da manhã e outro das 12:00 às 15:00. É possível que o grupo que aparece à 00:00 corresponda aos taxistas noturnos de Roma que ainda não encerraram a jornada de trabalho.

A Tabela 2 apresenta medidas estatísticas das distâncias diárias percorridas pelos táxis em Roma, e as velocidades calculadas após a aplicação do TTP e após a aplicação do TVA. Considerando todos os dias da base de dados de Roma, a maior distância diária obtida foi de 486,87 quilômetros, e o percentil 99 foi 355 quilômetros (27% menor). Estas distâncias são cerca de metade das calculadas para San Francisco, sugerindo que os táxis de San Francisco eram utilizados por mais de um taxista em mais de um turno do dia.

Tabela 2. Distâncias diárias e velocidades em Roma

	Distâncias Diárias (km)		Velocidades (km/h)	
Medidas	após TTP	após TVA	após TTP	após TVA
1º Quartil	283,19	124,38	0	0
Mediana	412,88	166,08	0	0
Média	480,67	171,19	34,34	15,86
3º Quartil	607,16	212,76	23,10	22,22
Percentil 90	860,65	262,47	56,82	50,77
Percentil 99	1.485,30	355,29	517,64	136,37
Maior	3.710,16	486,87	236.555,19	149,99
Coef. de variação	0,62	0,40	7,80	2,67

As análises de distâncias diárias dos táxis nos permitem visualizar e comparar padrões do funcionamento dos táxis em ambas as cidades. Tais resultados foram possíveis devido ao TVA ter tratado medidas anômalas obtidas quando o táxi estava movimentando.

6. Conclusão

Tratar e analisar grandes quantidades de dados é uma tarefa cada vez mais comum. Na etapa de pré-processamento, os dados são verificados para identificar inconsistências, possíveis erros e incompletude. Bases de dados que combinam informações de tempo e espaço podem tornar a análise ainda mais complexa devido à natureza contínua destas grandezas, associada às limitações das medições e erros. Estes erros podem ocorrer nos dispositivos e ferramentas utilizadas para obter e processar estes dados.

Neste trabalho, apresentamos um algoritmo para detecção e correção de velocidades anômalas, inconsistentes com as vias de tráfego, bem como a análise do impacto das correções realizadas. Estas anomalias ocorreram possivelmente devido a erros nos *timestamps* registrados nas bases de dados, e devido a erros na localização GPS dos táxis, nos mapas utilizados ou na *API* utilizada para o cálculo das distâncias. Vencidas estas etapas, analisamos padrões temporais do funcionamento dos táxis ao longo do dia com base nas distâncias tratadas. Estes padrões evidenciam horários de menor funcionamento do serviço e períodos do dia em que os taxistas iniciam suas atividades. O algoritmo e as análises realizads podem ser aplicadas a dados de localização de qualquer tipo de veículo.

Muitas análises podem ser feitas em trabalhos futuros. Por exemplo, podemos definir limiares de velocidades diferentes para os diferentes tipos de via; realizar o tratamento de velocidades com base na aceleração do veículo; e estimar o faturamento e lucro dos taxistas no decorrer do dia utilizando as distâncias após o tratamento. Além disto, podemos também confrontar os dados obtidos com a legislação local acerca do serviço de táxi para identificar como os limites legais são refletidos nas bases de dados.

7. Agradecimentos

Os autores agradecem ao CEFET-MG e aos financiadores dos projetos INCT InWeB (MCT/CNPq 573871/2008-6) e MASWeb (FAPEMIG/PRONEX APQ-01400-14).

Referências

Alvarenga, D., da Cunha, F. D., Viana, A. C., Mini, R. A., and Loureiro, A. A. (2016). Classificando comportamentos sociais em redes veiculares. In *XXXIV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*. SBC.

Bracciale, L., Bonola, M., Loreti, P., Bianchi, G., Amici, R., and Rabuffi, A. (2014). CRAWDAD dataset roma/taxi (v. 2014-07-17). Downloaded from http://crawdad.org/roma/taxi/20140717.

Ganti, R., Srivatsa, M., Ranganathan, A., and Han, J. (2013). Inferring Human Mobility Patterns from Taxicab Location Traces. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 459–468. ACM.

Gibbons, J. D. and Chakraborti, S. (2003). *Nonparametric Statistical Inference*. Marcel Dekker, New York.

Jones, K., Liu, L., and Alizadeh-Shabdiz, F. (2007). Improving Wireless Positioning with Look-Ahead Map-Matching. In *Fourth Annual International Conference on Mobile and Ubiquitous Systems: Networking & Services (MobiQuitous)*, pages 1–8. IEEE.

Júnior, A. M. S., Sousa, M. L., Xavier, F. Z., Xavier, W. Z., Almeida, J. M., Ziviani, A., Rangel, F., Avila, C., and Marques-Neto, H. T. (2016). Caracterização do Serviço de Táxi a partir de Corridas Solicitadas por um Aplicativo de Smartphone. In *XXXIV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*. SBC.

Kozievitch, N. P., Gadda, T. M. C., Fonseca, K. V. O., Rosa, M. O., Gomes-Jr, L. C., and Akbar, M. (2016). Exploratory Analysis of Public Transportation Data in Curitiba. In *43o. Seminário Integrado de Software e Hardware (SEMISH)*. SBC.

Monteiro, C. M. (2016). Padrões de Mobilidade Urbana em Serviços de Táxi. Mestrado em Modelagem Matemática e Computacional, Centro Federal de Educação Tecnológica de Minas Gerais – CEFET-MG, Belo Horizonte.

Monteiro, C. M., Silva, F. R., and Murta, C. D. (2016). Análise de Padrões Espaciais e Temporais da Mobilidade de Táxis em San Francisco e Roma. In *43o. Seminário Integrado de Software e Hardware (SEMISH)*. SBC.

Oliveira, A., Souza, M., de A. Pereira, M., Reis, F. A. L., Almeida, P. E. M., Silva, E. J., and Crepalde, D. S. (2015). Optimization of Taxi Cabs Assignment in Geographical Location-based Systems. In *XVI Brazilian Symposium on GeoInformatics*, pages 92–104. SBC.

Piorkowski, M., Sarafijanovic-Djukic, N., and Grossglauser, M. (2009). CRAWDAD dataset epfl/mobility (v. 2009-02-24). Downloaded from http://crawdad.org/epfl/mobility/20090224.

Rossi, L., Walker, J., and Musolesi, M. (2015). Spatio-temporal techniques for user identification by means of GPS mobility data. *EPJ Data Science*, 4(1):1–16.

Valero, B., Luis, J., Julián, A., Belén, A., Villén, G., and Natalia (2014). *GNSS. GPS: Fundamentos y Aplicaciones en Geomática*. Editorial de la Universidad Politécnica de Valencia, Valencia.