# T-MAPS: Modelo de Descrição do Cenário de Trânsito Baseado no Twitter

**Bruno P. Santos[1], Paulo H. L. Rettore[1], Heitor S. Ramos[2], Luiz F. M. Vieira[1], Antonio A. F. Loureiro[1]**

[1] Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte – MG – Brasil

[2]LACCAN/LCCV/Instituto de Computação
Universidade Federal de Alagoas (UFAL) – Maceió – AL – Brasil

[1]{bruno.ps, rettore, lfvieira, loureiro}@dcc.ufmg.br

[2]heitor@ic.ufal.br

***Abstract.*** *Understanding urban mobility, specifically transit mobility, has been focous of many research and investiments. However, obtain free access to real traffic data it is still limited, because the data is usually private and install sensors on transport network is expensive. In this paper, we caracterize the use of a Location Based Social Media (LBSM) plataforms to obtain a cost efficient transit network sketch. Also, we propose Twitter MAPS (T-MAPS) a description model of the transit scenario based on tweets. T-MAPS aim to improve todays' navegation systems and urban planning with information from LBSMs. In the avaluation, T-MAPS suggested routes with, in average, 62% of similarity with Google Directions' routes. Also, in 25% of the routes avaluated, the similarity reached 87% to 100%. These results indicates significant relationship between T-MAPS and Google Directions routes, even T-MAPS using solely tweets.*

***Resumo.*** *A compreensão da mobilidade urbana, em específico do trânsito, tem sido alvo de estudos e investimentos. Contudo, a obtenção de dados que representam o cenário do trânsito apresenta limitações, seja de acesso aos dados ou pelo elevado custo em implantar sensores na malha viária. Neste artigo, foi caracterizado o uso uma plataforma LBSM (Twitter), a fim de se obter uma descrição, de baixo custo, do trânsito. Também foi proposto o Twitter MAPS (T-MAPS), um modelo de descrição do cenário de trânsito baseado em tweets, com o objetivo de enriquecer as navegações e planejamento urbano com informações sobre regiões. Na avaliação, as sugestões de rotas do T-MAPS mostraram-se, em média, 62% similares às do Google Directions e, em 25% dos casos, a similaridade observada foi entre 87% e 100%, o que indica significativa relação entre as abordagens, mesmo T-MAPS usando somente tweets.*

## 1. Introdução

A economia e qualidade de vida em uma cidade é, em parte, reflexo da mobilidade que ela oferece. A infraestrutura de transporte local deve possibilitar o deslocamento das pessoas, o recebimento de insumos e escoamento das produções das empresas de forma eficiente.

Isso implica na necessidade de constante planejamento, gerência e manutenção dos sistemas de transporte. Nesse aspecto, a compreensão da mobilidade urbana, em específico da mobilidade no trânsito, tem despertado o interesse dos órgãos governamentais de planejamento e da academia. Contudo, para compreender o cenário de trânsito, é fundamental a obtenção e acesso a dados, tais como contagem de veículos passando por *loops* indutivos ou câmeras de trânsito, *traces* de usuários nas vias de transporte, matrizes de origem e destino de determinada região, dentre outros. O acesso a estes dados representa desafios na compreensão do cenário de trânsito de uma região, pois grande parte dos dados são controlados por entidades privadas ou governamentais e, geralmente, são inacessíveis em parte ou em sua totalidade, seja por questões de mercado ou mesmo privacidade. Como alternativa de baixo custo ao cenário atual para obtenção de dados, surge a oportunidade de usar as mídias sociais baseadas em localização – *Location Based Social Media (LBSM)*) como, por exemplo, *Twitter* e *Foursquare*, com o objetivo de suprir a deficiência de dados sobre um contexto específico.

As LBSMs são amplamente utilizadas por usuários que fornecem os dados, sejam eles corporativos ou não. Segundo o *Twitter*, cerca de 313 milhões de usuários estão ativos todos os meses (dados de junho de 2016) em sua rede. Isto pode gerar diversas oportunidades em busca de melhor compreender uma determinada situação do trânsito, por exemplo. Contudo, os dados geralmente apresentam particularidades, inerentes ao uso pelas pessoas, que podem acarretar em grandes desafios como, por exemplo, dados imprecisos, enviesados e inconsistentes.

Neste trabalho, um estudo piloto foi realizado para melhor entender o relacionamento entre o cenário real do trânsito e os dados de uma plataforma LBSMs, o Twitter. A partir desse estudo, os detalhes das características dos dados foram levantadas e classificadas. Após essa caracterização, foi desenvolvido o Twitter MAPS (T-MAPS), um modelo de descrição do cenário de trânsito baseado no Twitter, com o objetivo de enriquecer o contexto de navegação atual, vinculando os *tweets* às rotas geradas por plataformas já consolidadas como, por exemplo, o *Google Directions*. O modelo é apresentado de forma generalista e pode ser aplicado a qualquer região. No entanto, neste trabalho os dados e, consequentemente, os resultados são provenientes do estado de Nova Iorque nos Estados Unidos.

As principais contribuições realizadas neste trabalho são:

- Caracterização de redes LBSM, em especial do Twitter, como uma fonte de dados para melhor entender e descrever o cenário de trânsito.
- Proposição do T-MAPS como um modelo de descrição cenário de trânsito baseado em dados extraídos do Twitter.

A motivação deste trabalho vem do desejo em ampliar o conhecimento do trânsito em uma dada região, usando dados obtidos de LBSMs, com o objetivo de proporcionar uma experiência mais descritiva desse cenário, pouco explorado ainda na literatura. A indicação do sentimento de um trajeto, a intensidade de acidentes e descrições mais detalhadas de ocorrências são exemplos de informações que podem enriquecer a experiência do usuário da infraestrutura de transporte e melhor auxiliá-lo nas decisões relacionadas a mobilidade urbana.

O restante deste trabalho está organizado como descrito a seguir. A Seção 2 apresenta as características da coleta de dados do *Twitter*. A Seção 3 descreve os problemas

**Tabela 1. Exemplo de contas**

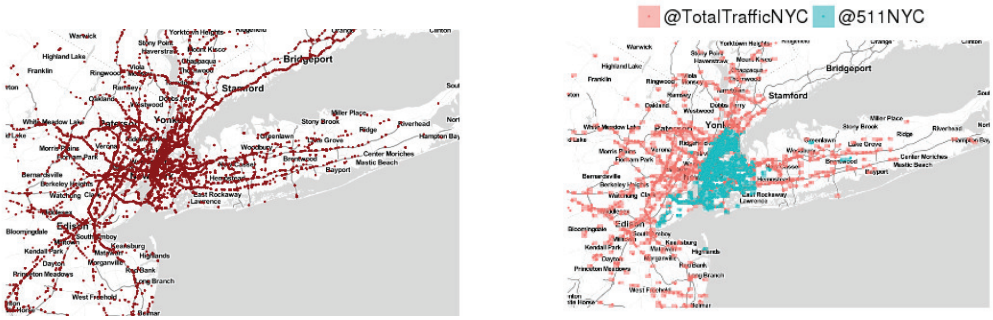
| Nome da Conta | # *Tweets* |
|---|---|
| @511NYC | 126925 |
| @TotalTrafficNYC | 20267 |
| @WazeTrafficNYC | 7850 |
| @Traffic4NY | 3789 |
| ... | ... |
| @NYC_DOT | 3680 |
| Total das 21 contas: | 355K |


**Figura 2. Exemplo de *Tweet***

existentes nesses dados, os quais devem ser tratados para serem utilizados. A Seção 4 descreve os passos da modelagem T-MAPS, que é o resultado da utilização desses dados. A avaliação da modelagem é realizada na Seção 5. Em seguida a Seção 6 apresenta os trabalhos relacionados e, por fim, a Seção 7 apresenta a conclusão e trabalhos futuros.

## 2. Coleta de Dados

O processo de coleta dos dados do Twitter foi conduzido utilizando a REST Aplication Program Interface (API)[1], seguindo os termos de limites da plataforma. Foram selecionados, manualmente, 21 contas especializadas em notificar o trânsito em todo o estado de Nova Iorque por meio de *tweets*. Essas contas são dedicadas apenas à publicação de eventos de trânsito através de *tweets* e são mantidas por diferentes organizações, incluindo órgãos governamentais. No período de tempo amostrado, às 21 contas reportaram um montante de aproximadamente 355 mil *tweets*. A quantidade de dados que apresentam etiquetas de geolocalização ultrapassa 307 mil e os dados que estão posicionados dentro da região de Manhattan é de aproximadamente 38 mil. Um exemplo de *tweet* publicado por uma das contas especialistas em trânsito é apresentado na Figura 2. É possível notar a variedade de informações disponibilizadas neste *tweet*; tem-se desde a descrição do evento ocorrido (um acidente) até horário, informações sobre congestionamento, direção e localização precisa. É importante notar que as contas de usuários comuns não foram consideradas devido ao maior viés de seus *tweets* em relação ao trânsito.




(a) *Tweets* na cidade de Nova Iorque (b) Cobertura espacial de duas contas

**Figura 3. Cobertura de espacial**

[1]https://dev.twitter.com/rest/public

A Figura 3 apresenta a cobertura espacial dos dados coletados na cidade de Nova Iorque e regiões adjacentes e, em específico, a Figura 3(a) destaca a cobertura de *tweets* com geolocalização coletados de todas as 21 contas especialista. Ao aproximar a imagem é possível notar que os *tweets* estão associados às vias na malha de transporte da cidade. A Figura 3(b) mostra a cobertura espacial de duas contas especialistas, @TotalTrafficNYC e @511NYC, as quais possuem maior número de publicações. O maior raio de cobertura é apresentado pela conta @TotalTrafficNYC, publicando *tweets* tanto no perímetro da cidade de Nova Iorque quanto nas regiões adjacentes, ao passo que a conta @511NYC limita-se à cidade. Esses dados apresentam a informação de que diferentes contas possuem diferentes áreas de cobertura. Cientes disto, é possível selecionar contas que se complementam aumentando, assim, a cobertura espacial de uma região de interesse.

Entender a cobertura temporal é outro ponto importante ao manipular dados de LBSMs. A Figura 4 mostra como três contas especialistas se comportam ao longo de uma semana. A conta @NYC_DOT está associada ao departamento de transporte da cidade de Nova Iorque e apresenta um típico comportamento de trabalho em horário comercial, não publicando *tweets* durante os finais de semana. Já @TotalTrafficNYC e @511NYC apresentam comportamento constante durante as semanas, mas variam as taxas de atividade ao longo de um dia, como mostrado na Figura 4 mais à direita. Essas informações também podem ser uteis para combinar contas e aumentar a cobertura temporal.

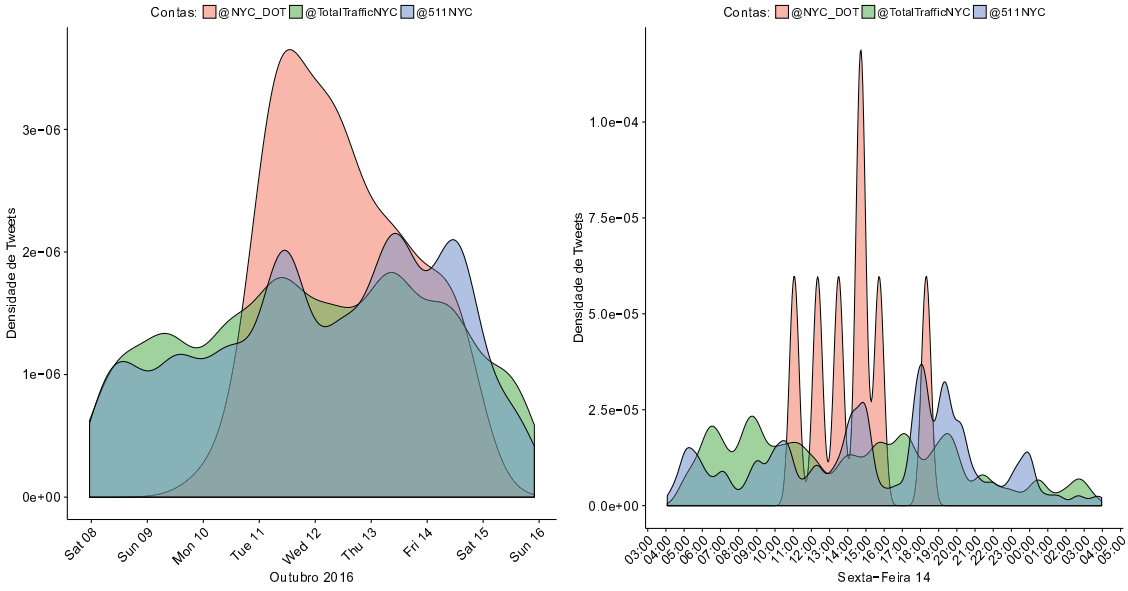



**Figura 4. Cobertura temporal de três contas especialistas**

Para complementar os pontos levantados anteriormente, a Figura 5 apresenta uma visão espacial e temporal dos dados coletados na Sexta-Feira, 14 de Outubro de 2016. Os dados foram agregados em faixas de uma hora, sendo possível observar que nas primeiras horas ocorreram poucos *tweets* na cidade de Nova Iorque. Por outro lado, é perceptível o aumento das atividades após as 6 horas da manhã, momento em que a população tende a iniciar suas atividades. Após às 12 horas, também é possível notar o aumento na quantidade de atividade, em especial, nos horários de 15 e 18 horas e, logo após esse horário, existe uma tendência de diminuição do número de *tweets* reportados por hora, até voltar

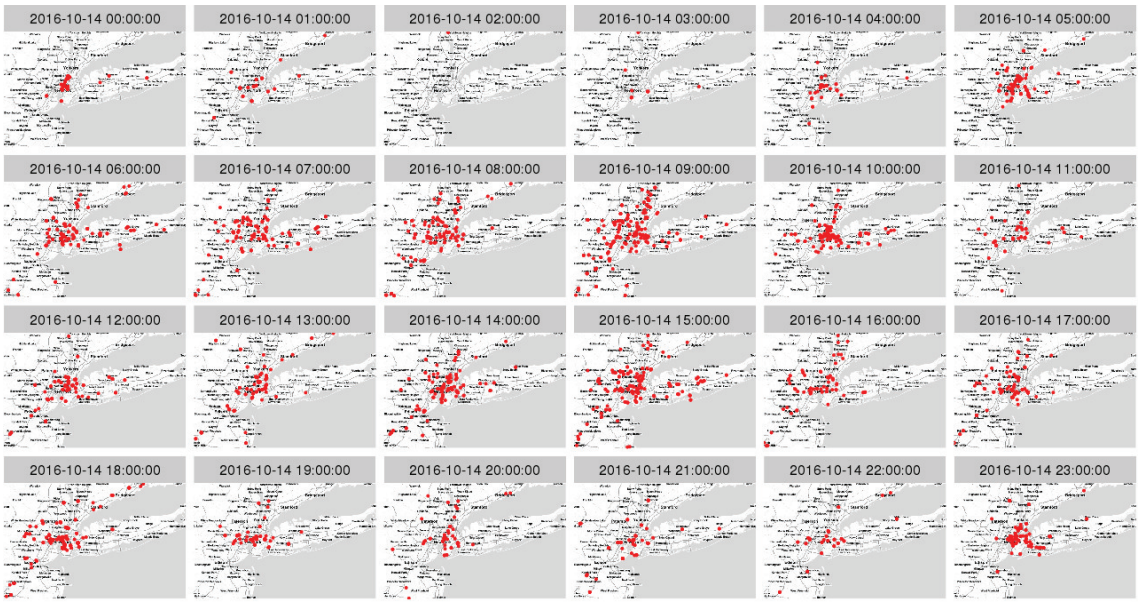



**Figura 5. Cobertura espaçotemporal**

a aumentar às 23 horas. Isso se deve ao fato do dia da semana analisado, pois supõe-se que esse aumento das atividades, especialmente concentrada em Manhattan, esteja relacionada ao início de atividades noturnas, típicas do início do final de semana.

## 2.1. Correlação entre *tweets* e trânsito

Identificar o quão bem relacionados estão os dados de LBSMs com o trânsito pode revelar a capacidade que tais dados têm de representar o cenário do trânsito. No estudo aqui apresentado, essa informação de relacionamento pode dizer se os *tweets* são relevantes para descrever, de algum modo, o cenário do trânsito no tempo. Por esse motivo, esta seção visa responder à seguinte pergunta: *os dados coletados do Twitter apresentam correlação com dados reais de trânsito?*

Para responder a esta pergunta é necessário obter dados reais do trânsito, tais como contagem de veículos passando por *loops* indutivos ou câmeras de trânsito, rastro de GPS de usuários na rede de transporte, matrizes de origem e destino de uma região, dentre outras. Esses dados serão considerados como dados tradicionais, os quais permitem estudar as interações entre demanda (motoristas, veículos, pedestres, ciclistas) e a infraestrutura (ruas, rodovias, sinais, dispositivos de controle, etc.). Essas interações ajudam a entender e desenvolver malhas de transporte mais eficientes, que otimizam o movimento do tráfego e reduzem congestionamentos [Bazzan and Klügl 2013].

Os dados tradicionais de trânsito têm por finalidade capturar três principais variáveis: velocidade, densidade e fluxo. Essas variáveis permitem visualizar o comportamento do trânsito [Kung et al. 2014, Zheng et al. 2008, Bazzan and Klügl 2013]. No entanto, os dados tradicionais isoladamente não capturam a experiência dos usuários que utilizam a malha de transporte; o sentimento que se tem a respeito das regiões; ou eventos com maior teor semântico como a gravidade de um acidente. Além disso, acessar dados tradicionais pode não ser uma tarefa fácil, já que muitos deles são privados (corporativos ou governamentais) como, por exemplo, trajetos de GPS de *smartphones*, trajetos de GPS

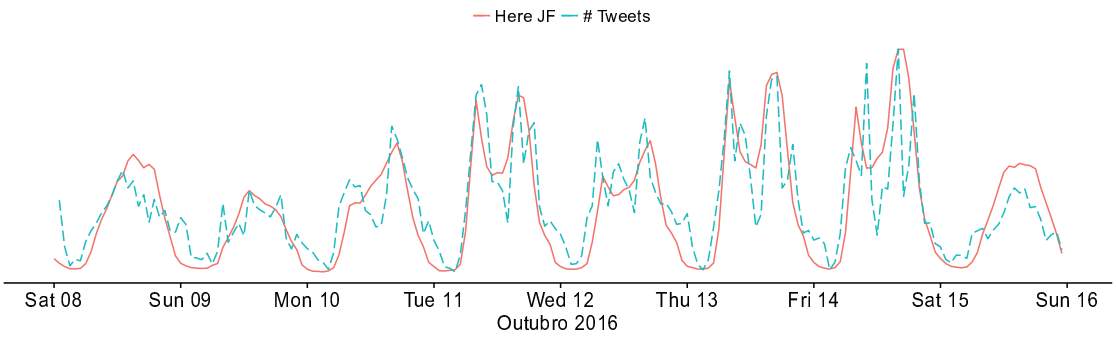



**Figura 6. Número de ocorrências de *tweets* e o fator de congestionamento provido pela empresa *HERE*. A correlação de Spearman para as séries é de 0.81**

veicular (ex: táxi) e *loops* indutivos.

Grandes empresas ou agências de transporte governamentais geralmente realizam processos de fusão sobre os dados observados. Esses processos têm por objetivo enriquecer o conhecimento sobre o estado atual do trânsito e facilitar sua visualização. Nessa direção algumas empresas como *HERE*, TomTom e *Google* permitem o acesso, *parcial*, aos seus mapas de trânsito que contêm o resultado do processo de fusão sobre os dados. Por exemplo, a empresa *HERE* fornece uma API de acesso ao dado *Jam Factor (JF)*[2] (fusão dos dados tradicionais e outros), o qual representa o nível de congestionamento das vias em que a empresa possui dados, ao passo que *Google* e *TomTom* não oferecem acesso direto a essa informação. Por esse motivo, o JF será utilizado no comparativo de relacionamento entre *tweets* e o contexto do trânsito.

A Figura 6 apresenta duas curvas em série temporal entre os dias 8 e 16 de Outubro de 2016. A curva tracejada representa a contagem de *tweets* ao longo do tempo, enquanto a curva sólida representa o fator de congestionamento fornecido pela API HERE (JF). As curvas foram normalizadas e estão em uma escala entre 0 e 1. É possível perceber o forte padrão periódico nessas séries de dados. Os dias da semana possuem maior contagem de *tweets* e fator de congestionamento quando comparado aos finais de semana. Esse padrão se repete em conjunto com dois pontos de maior contagem e JF.

O coeficiente de correlação de Spearman foi utilizado como medida de relacionamento entre os *tweets* e o fator de congestionamento. Essa correlação é mais adequada nesse cenário do que a tradicional correlação de Pearson, pelo fato das séries de dados não apresentarem comportamento linear. Tal como a correlação de Pearson, a de Spearman apresenta valores entre $-1$ e $+1$ para correlações perfeitas. O coeficiente resultante entre as duas séries de dados foi de *0.81*. A interpretação dessa correlação indica que os dados estão altamente relacionados e que, por ser uma correlação positiva, em momentos nos quais o fator de congestionamento cresce o número de *tweets* também tende a crescer.

## 3. Aspectos dos Dados

Observada a alta correlação entre os dados do *tweets* e o cenário do trânsito, esta seção apresenta alguns aspectos que os dados do Twitter e de LBSMs podem apresentar. Os dados geralmente apresentam particularidades que podem acarretar em dificuldades no

[2] https://developer.here.com/rest-apis/documentation/traffic/topics/quick-start.html

uso para determinar o cenário do trânsito. Os dados serão classificados em cinco aspectos e, em seguida, serão discutidos os problemas e potenciais soluções, sempre que possível.

### 3.1. Imprecisão

Dados providos pelos *tweets*, quando fazem referência a informações de trânsito, podem apresentar alguma imprecisão. Isso é revelado quando os dados apresentam ao menos uma das seguintes características: dados incompletos; vagos ou granularidade que afeta sua interpretação. É esperado que imprecisões apareçam nos dados causadas pela inerente heterogeneidade das fontes e do alto grau de liberdade da entrada de dados. A seguir, é apresentado um exemplo de *tweet* não geolocalizado que será utilizado para ilustrar os aspectos de imprecisão descritos:

*"Agora 8:00AM um acidente na Av. Antônio Carlos #BH #tráfegoRuim #asustado"*.

- **Incompleto:** é comum que *tweets* não apresentem informações completas do evento ocorrido, principalmente pela limitação de espaço do texto. No *tweet* do exemplo, é possível obter informações tais como sentimento do usuário, condição do tráfego (para aquele usuário) e o horário. Entretanto, os dados diretamente acessíveis são insuficientes (e.g., inexistência de etiqueta de geo-localização e descrição textual) para derivar facilmente a localização do evento. Uma possível forma de solução é a aplicação de técnicas de aprendizado de máquina ou *record-linkage*.
- **Vago:** está relacionado ao grau de clareza da informação sobre o evento e seu contexto. No *tweet* do exemplo, o dado está vago, pois ao extrair as informações do texto, pouco é dito sobre o local exato do incidente (a Av. Antônio Carlos, possui mais de 8 km de extensão). Neste caso, o *tweet* é vago em relação à localização do evento.
- **Granularidade:** varia de baixa a alta granularidade. Quando são de baixa granularidade, os dados contêm informações suficientes para descrever precisamente os seguintes itens: localização do evento, sentido, gravidade e outras informações de interesse. Caso contrário, são considerados de alta granularidade e apresentam uma visão macro dos eventos.

### 3.2. Viés dos usuários

Os dados obtidos e analisados pelas LBSMs podem conter vieses, uma vez que, as informações sobre trânsito e tráfego podem ser reportadas por qualquer usuário. Por exemplo, os congestionamentos podem ser percebidos de diferentes maneiras. Suponha o seguinte cenário onde o usuário de uma pequena cidade está no trânsito de uma grande metrópole: nesse caso, ele pode perceber o congestionamento de forma diferente de um usuário que reside nessa metrópole. Consequentemente, é introduzido ao dado a percepção do usuário, gerando problemas como inconsistência e conflitos.

Até mesmo usuários especialistas[3] podem introduzir vieses. Usuários dessa classe podem, por exemplo, manter um público alvo específico ou ainda apresentar maior ou menor intenção de publicar de informações de determinado local em detrimento de outros.

[3]Consideram-se como usuários especialistas as contas que têm como propósito reportar apenas informações sobre o cenário de trânsito.

No conjunto de dados explorados nesse trabalho, foram selecionados manualmente os usuários que representam contas corporativas, ou seja, entidades como jornais ou departamento de transporte local. Desse modo, apesar da natureza diversa dessas contas, os dados podem seguir vieses inerentes aos interesses dessas contas.

### 3.3. Atribuição espacial e temporal

A atribuição espacial e temporal são os pontos mais críticos dos aspectos dos dados providos por LBSMs. A geo-localização e o aspecto temporal dos dados permitem caracterizar uma dada região e um instante ou intervalo de tempo. Por outro lado, os dados que não apresentam esses aspectos podem não fazer sentido para a construção do cenário atual do trânsito. A seguir, serão descritos os principais detalhes da atribuição espacial e temporal no processo de extração dessas informações de uma LBSMs:

- **Espacial:** atribuir uma localização ao dado é fundamental para entender o contexto que envolve a informação. Contudo, derivar essas informações, mesmo estando presente nos dados ou meta-dados dos *tweets*, não é uma tarefa trivial. Suponha, por exemplo, que a informação da posição no espaço esteja incorporada ao texto de um *tweet*. Portanto, extrair tal informação requer uma estrutura que possibilite a identificação desses dados no texto. Entretanto, o texto de um *tweet* é inerentemente desestruturado, permite poucos caracteres e existem formas diferentes de se escrever o texto, o que resulta, muitas vezes, em subjetividade na interpretação das informações (e.g., "Avenida" e "Av.", "R. Nome" e "R." significa Rua ou Rodovia?). Desse modo, técnicas de *Natural Language Processing (NLP)* [Liu et al. 2011, Li and Sun 2014] têm obtido resultados interessantes na extração de tais informações. Em [Li and Sun 2014], os autores apresentam uma técnica baseada em NLP, chamada *Petar*, para extrair Point of Interests (POIs) dos textos dos *tweets*.
  Disponibilidade de informação é outro ponto que afeta a atribuição espacial. Espera-se que algumas regiões tenham maior cobertura espacial que outras por diversos fatores. Por exemplo, grandes cidades tendem a apresentar maior cobertura espacial do que cidades pequenas.
- **Temporal:** associar um aspecto temporal (*timestamp*) ao dado publicado é de suma importância para entender o cenário passado, atual e, possivelmente, futuro da malha de transporte. Em geral, as plataformas LBSM associam uma marca de tempo no momento em que a informação é reportada. Entretanto, essa marcação não necessariamente é a mesma do momento em que o evento ocorreu. Assim, surgem alguns questionamentos em relação a esse aspecto tais como: *Qual é a validade de um dado publicado por um usuário de LBSM? Como se caracteriza o atraso entre o evento e o surgimento da informação nas plataformas LBSM?*

### 3.4. Inconsistências

Nesta seção, são abordadas questões que podem gerar inconsistências no uso de dados de plataformas LBSM. A noção de inconsistência diz respeito aos aspectos de conflito e desordem dos dados. Uma visão semelhante é apresentada em [Rettore et al. 2016].

- **Conflito:** os dados de LBSMs são conflitantes quando duas ou mais fontes de informação divergem a respeito de um evento. Por exemplo, suponha que dois usuários do Twitter publiquem *tweets* a respeito do mesmo evento, como

um possível incidente. Um dos usuários relata que nada de grave ocorreu e o fluxo de trânsito está bom, enquanto o outro usuário reporta um grave acidente que gerou impacto negativo no trânsito. Nesse caso, somente com essas duas informações não se pode dizer se, de fato, ocorreu um acidente e nem as consequências do evento. Soluções baseadas em teoria da evidência (Dempster-Shafer) têm ganhado notoriedade em reduzir divergências de dados conflitantes [Zadeh 1984, Florea et al. 2009]. Outra abordagem, mais simplista, é empregar diferentes pesos às informações providas por diferentes contas de usuários, por exemplo, usuários comuns recebem peso $X$ enquanto usuários especialistas recebem peso $Y$ e, assim, aplicar regras para decidir sobre as informações.

- **Fora de ordem:** a liberdade oferecida pelas plataformas LBSMs permitem aos usuários inserir dados fora de ordem, no tempo. Esses dados aparecem como inconsistentes para os sistemas que venham a utilizá-los. Esse aspecto se relaciona como o aspecto temporal dos dados, e a principal questão que surge é como empregá-los de forma adequada. Uma possível solução é descartar os dados que estão fora de ordem. Entretanto, tal solução implica em perda de informação e, consequentemente em degradação da cobertura espacial e temporal do sistema. Uma segunda alternativa é armazenar todos os dados obtidos e ordená-los, necessitando mais recursos computacionais como processamento e armazenamento.

## 4. Processo de modelagem do Twitter MAPS (T-MAPS)

Esta seção apresenta uma modelagem que permite utilizar os dados coletados (veja a Seção 2) do Twitter como uma representação do cenário de trânsito. A modelagem proposta foi desenvolvida de forma flexível, com o objetivo de possibilitar a agregação de dados de outras plataformas LBSMs. O processo de modelagem é descrito a seguir:

- ***Passo 1 – Aquisição de informações:*** essa etapa consiste em segmentar a área de interesse e obter dados das plataformas LBSMs. A segmentação é realizada devido as inerentes lacunas na cobertura espacial dos dados, o propósito desta segmentação é viabilizar uma visão macro do trânsito local a partir da associação dos dados com o mapa.
- ***Passo 2 – Filtragem e fusão de dados:*** cria-se um grafo ponderado e dirigido para representar a segmentação realizada. O grafo é definido como $G = (V, A)$, tal que, $V(G)$ é o conjunto de vértices que representam as divisões da área e $A(G)$ é o conjunto de arestas entre quaisquer dois vértices de $G$, desde que as áreas que eles representam sejam adjacentes no mapa real.
- ***Passo 3 – Métricas de custo:*** esse passo consiste em atribuir pesos às arestas direcionadas da seguinte forma: $A(x, y) \leftarrow \text{peso}(y)$. Esse peso representa a descrição do cenário do trânsito.

A Figura 7 ilustra as etapas da modelagem proposta e as diferentes formas de ponderar o grafo. Iniciando com o mapa de regiões de Nova Iorque, obtém-se o grafo $G = (V, A)$, sendo $V(G)$ as regiões e $A(G)$ as adjacências. Em seguida, a camada de dados do *Twitter* é aplicada ao grafo com o objetivo de posteriormente acrescentar informações de mídias sociais como peso às arestas do grafo. O modo utilizado para ponderar o grafo $G$ consiste em $A(v, w) \leftarrow \text{peso}(w)$, onde $\text{peso}(w)$ pode seguir várias estratégias. Abaixo são listadas algumas dessas estratégias:

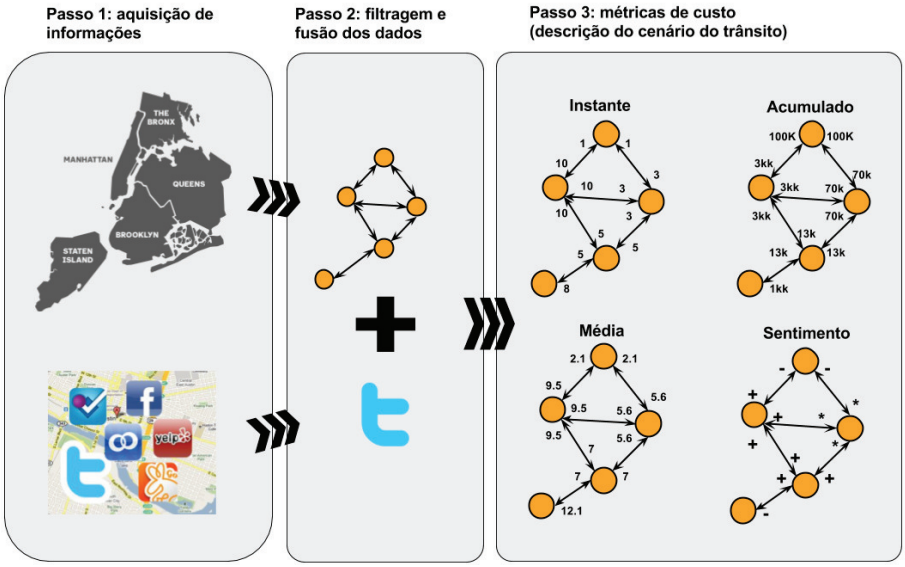



**Figura 7. Processo de modelagem**

- **Instante:** abordagem que considera os *tweets* de um dado dia e instante no tempo, correspondendo ao agregado de uma hora fechada. Esta estratégia pode ser considerada como uma caricatura do trânsito naquele momento.
- **Acumulado:** consiste na agregação dos *tweets* ao longo do tempo, desde que correspondam ao mesmo dia da semana e hora fechada. Nesse caso, leva-se em conta o histórico daquela região para ponderar as arestas do grafo.
- **Média:** apropria-se da mesma abordagem do *Acumulado*, contudo os *tweets* representam a média das ocorrências ao longo do tempo, dia da semana e hora fechada. Essa estratégia pode ser utilizada como cenário padrão médio do trânsito e sempre que novos dados entram no sistema, pode-se comparar com a média histórica e, então inferir sobre o trânsito.
- **Sentimento:** o texto do *tweet* é analisado, com o objetivo de associar um sentimento a esse texto. Com essa atribuição, pode-se considerar o sentimento (e.g., positivo, negativo ou neutro) que mais influencia cada região.

A modelagem desenvolvida, apresentada como T-MAPS, busca descrever o cenário de trânsito baseando-se nos *tweets* de uma dada região. Apoiado nesse processo, o T-MAPS busca prover serviços que auxiliem os usuários a compreender o trânsito de maneira mais abrangente. Dentre os serviços propostos pelo T-MAPS, destacam-se: prover uma visão macro do trânsito, indicar o sentimento das regiões ao longo do tempo, sugerir rotas baseadas na frequência de ocorrência registradas nos *tweets*, sugerir rotas baseadas no sentimento das regiões e destacar rotas com elevado número de acidentes ou desastres. É importante salientar que, ferramentas como *Google/Microsoft/TomTom* exploram outros aspectos na sugestão das rotas de trânsito. Assim, os serviços do T-MAPS constituem uma alternativa para enriquecer os cenários que se têm do trânsito com informações publicadas em plataformas LBSMs.

## 5. Avaliação do T-MAPS

Nesta seção, são apresentadas avaliações do T-MAPS. Inicialmente o serviço de rotas do T-MAPS é avaliado utilizando o *Google Directions*, aqui considerada como a

representação mais fiel do cenário de trânsito. Posteriormente é apresentado o serviço de sentimento das rotas do T-MAPS, com o objetivo de descrever com mais detalhes os trajetos fornecidos pelo *Google Directions*.

Os resultados apresentados são referentes à segmentação da região de Manhattan em Nova Iorque. Após as análises de frequência de *tweets* em Manhattan, notou-se que a cobertura espacial e temporal dos dados viabiliza o uso das subdivisões do mapa. No nível de bairros (existem 29 bairros oficiais em Manhattan[4]), todos os bairros apresentam *tweets* nas faixas de tempo analisadas. Seguindo a modelagem descrita anteriormente, obtivemos grafos para cada estratégia de custo (veja Seção 4). Os grafos contêm 29 vértices (bairros) e arestas entre os bairros adjacentes.

### 5.1. Serviço de rotas T-MAPS

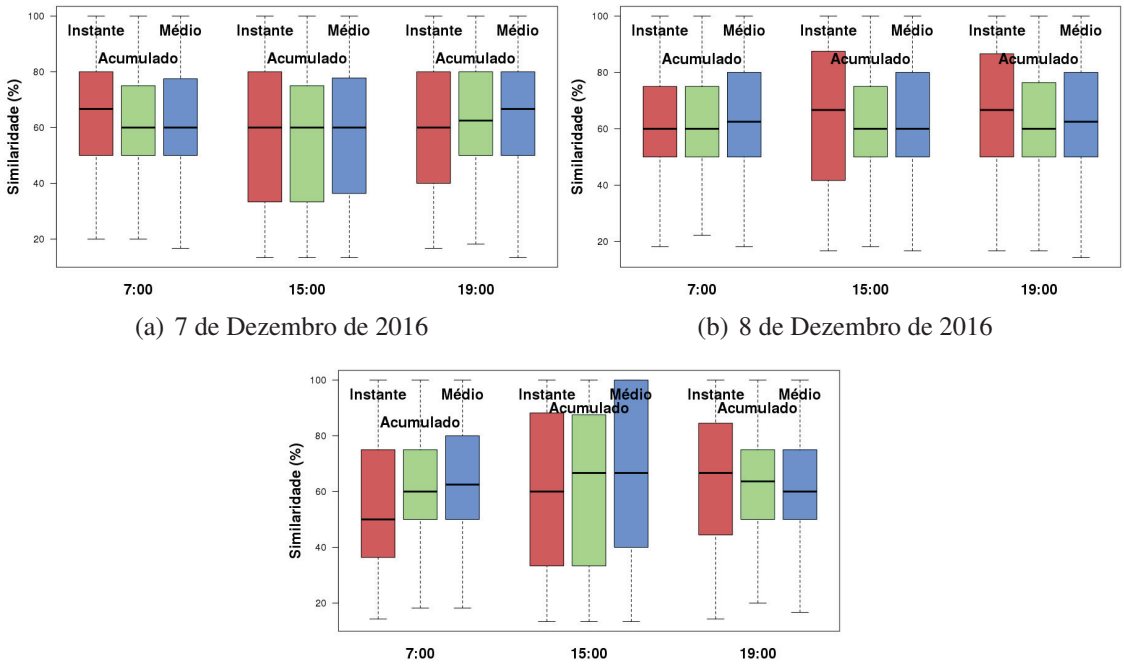
Este serviço do T-Maps sugere rotas considerando uma visão macro entre regiões no mapa, ou seja, o objetivo é dizer ao usuário que as regiões recomendadas possuem as melhores condições segundo a métrica aplicada. A avaliação deste serviço foi realizada comparando as regiões percorridas pelas rotas providas pelo T-MAPS[5] e *Google Directions*. A comparação entre os trajetos foi baseada na similaridade entre as regiões sugeridas, ou seja, buscou-se identificar o percentual de similaridades da abordagem T-MAPS em relação ao *Google Directions*. Foram geradas 812 rotas a partir da combinação de cada um dos 29 bairros (como origens e destinos) de Manhattan, exceto aqueles em que a origem e destino são os mesmos. Além disso, rotas de $A \rightarrow B$ e de $B \rightarrow A$ também são consideradas. Essas rotas foram coletas do T-MAPS e *Google Directions* em três momentos do dia (7, 15 e 19 horas) escolhidos propositalmente, pois, geralmente, são os horários onde as ocorrências de *tweets* e congestionamentos são mais expressivos.

As similaridade é definida como o percentual de interseção das rotas recomendadas pelo T-MAPS e *Google Directions*. A Figura 8 apresenta uma amostra da avaliação de similaridade em três dias de coleta para três métricas de custo do T-MAPS, totalizando 21924 rotas analisadas. É possível notar que, a métrica *Instante* resulta na maior variação da mediana da taxa de similaridade, entre 50% e 66%, enquanto as métricas *Acumulado* e *Médio* apresentam variação entre aproximadamente 60% e 66%. Interpreta-se que na métrica *Instante*, ao menos a metade das rotas consideradas possuem 50% de similaridade com as rotas do *Google Directions*, ao passo que a métrica *Acumulado* e *Médio* apresentam ao menos 60% de similaridade. Também é importante ressaltar que, na média de todos os dias e horários avaliados, metade das rotas apresentaram ao menos 62% de similaridade entre T-MAPS e *Google Directions*. Além disso, uma porcentagem relevante de 25% das rotas apresentou grau de similaridade entre 87% e 100%. Finalmente, ressalta-se que 75% das rotas avaliadas apresentam grau de similaridade superior a 47%

Portanto, uma modelagem que descreve o cenário de trânsito, baseada exclusivamente nos *tweets*, fornece significativo grau de correspondência nos cenários avaliados, quando comparado com a ferramenta *Google Directions*, que por sua vez baseia-se em diversas fontes de dados. Em consequência desta similaridade, o T-MAPS pode contribuir com o enriquecimento do cenário de trânsito fornecendo contexto, baseado nos textos dos *tweets*. Um dos possíveis serviços de enriquecimento das rotas é a análise dos sentimentos

[4] `www1.nyc.gov/site/planning/data-maps/open-data/districts-download-metadata.page`

[5] Foi usado o algoritmo de Dijkstra sobre o grafo ponderado para computar o menor caminho.




(a) 7 de Dezembro de 2016

(b) 8 de Dezembro de 2016

(c) 9 de Dezembro de 2016

**Figura 8. Similaridade entre rotas sugeridas pelo T-MAPS e *Google Directions***

de um trajeto, o qual será explorado na seção seguinte.

## 5.2. Serviço de sentimento da rota

Esta seção descreve uma possível técnica para prover o serviço de sentimento de rota usando o T-MAPS. Para isso, realizou-se um pré-processamento sobre o texto de todos os dados coletados. Inicialmente foram removidas todas as palavras de parada (*Stop Words*), *links*, caracteres especiais e pontuação. Em seguida, executou-se um processo chamado *stemming* sobre o texto. O *stemming* remove sufixos das palavras como, por exemplo, *"Jaming"*, *"Jammed"* para que todas sejam consideradas como apenas *Jam*. Como resultado, gerou-se a nuvem de palavras exibida na Figura 9(a), em que o tamanho da palavra indica a maior frequência de uso nos *tweets*.

Para extrair o sentimento do texto dos *tweets*, foi utilizado o pacote de acesso livre *syuzhet*[6] da ferramenta R. Esse pacote contém algoritmos para extração de emoções e sentimentos de textos. A ideia dos algoritmos é utilizar dicionários contendo várias palavras e emoções/sentimentos associados a elas. A definição da emoção ou sentimento de um determinado texto depende do número de ocorrências dessas palavras no texto, ou seja, verificam-se as palavras mais comumente utilizadas e o grau de emoção/sentimento é atribuído ao texto.

A Figura 9(b) apresenta o resultado da análise de sentimento no mapa e uma rota sugerida pelo *Google Directions*. Com essas duas informações (rota e sentimento), o T-MAPS pode enriquecer o trajeto sugerido, indicando aos usuários o sentimento das regiões do trajeto ou até mesmo prover o roteamento baseado nestes sentimentos.

[6]`https://cran.r-project.org/web/packages/syuzhet/syuzhet.pdf`


(a) Palavras frequentes nos *tweets* (b) Manhattan dia 14 de Out

**Figura 9. Análise de sentimentos sobre os texto dos *tweets* coletados**

## 6. Trabalhos Relacionados

Diversos trabalhos com o intuito de analisar os sentimentos de determinada região, ou ainda, propor recomendações de locais baseadas em informações obtidas de LBSMs, são encontrados na literatura [Bertrand et al. 2013, Kim et al. 2014]. Contudo, de forma diferente dos trabalhos citados, o estudo e a modelagem aqui apresentados propõem associar dados de LBSMs, em especial do Twitter, às regiões do mapa de interesse e, então, enriquecer o contexto de navegação atual como, por exemplo, vinculando os sentimentos dos *tweets* às rotas geradas por plataformas tais como o *Google Maps*.

Na difícil tarefa de explorar o texto dos dados, alguns trabalhos têm obtido alguns avanços que podem aperfeiçoar o T-MAPS. Em [Ribeiro Jr et al. 2012], os autores apresentam um sistema para detectar e localizar eventos usando o texto de *tweets*. Em [Gong et al. 2015], é apresentado um sistema que identifica e sinaliza aos usuários pontos de congestionamento utilizando LBSMs. Em relação a estudos do comportamento, mobilidade, trajetórias e demandas de viagens diversos trabalhos têm recebido destaque [Yin and Du 2016, Kung et al. 2014, Cho et al. 2011, Zheng et al. 2008]. Esses trabalhos apresentam de um modo ou de outro uma complementação à abordagem apresentada neste trabalho, visando compreender melhor o cenário do trânsito como um todo.

## 7. Conclusão e Trabalhos Futuros

Este trabalho apresentou um estudo de caracterização e relacionamento entre dados de plataformas LBSMs e o cenário de trânsito. Além disso, foi apresentado o T-MAPS que usa um modelo baseado em grafo e dados do Twitter, com o objetivo de enriquecer o contexto de navegação na cidade, associando informações dos *tweets* às rotas geradas. Na avaliação dos serviços providos pelo T-MAPS, mostrou-se que as rotas sugeridas apresentam em média 62% de similaridade com as do *Google Directions* e em 25% das rotas avaliadas foram obtidos graus de similaridade entre 87% e 100%. Isso indica que o T-MAPS fornece significativa correspondência com as rotas do *Google Directions*, mesmo utilizando exclusivamente *tweets*. Também foi apresentado um serviço de enriquecimento das rotas sugeridas pelo *Google Directinos* baseado no sentimento das regiões, por sua vez extraídos do processamento dos textos dos *tweets*.

Finalmente, pretende-se como trabalhos futuros estender o T-MAPS para regiões com maiores dimensões, explorar os textos dos *tweets* a fim de se extrair mais informações sobre os eventos reportados. Além disso, pretendemos avançar em questões relacionadas à confiabilidade das fontes e validade dos dados.

## Referências

[Bazzan and Klügl 2013] Bazzan, A. L. and Klügl, F. (2013). Introduction to intelligent systems in traffic and transportation. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 7(3):1–137.

[Bertrand et al. 2013] Bertrand, K. Z., Bialik, M., Virdee, K., Gros, A., and Bar-Yam, Y. (2013). Sentiment in new york city: A high resolution spatial and temporal view. *arXiv preprint arXiv:1308.5010*.

[Cho et al. 2011] Cho, E., Myers, S. A., and Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks. In *17th ACM SIGKDD*, pages 1082–1090. ACM.

[Florea et al. 2009] Florea, M. C., Jousselme, A.-L., Bossé, É., and Grenier, D. (2009). Robust combination rules for evidence theory. *Information Fusion*, 10(2):183–197.

[Gong et al. 2015] Gong, Y., Deng, F., and Sinnott, R. O. (2015). Identification of (near) real-time traffic congestion in the cities of australia through twitter. In *ACM Understanding the City with Urban Informatics*.

[Kim et al. 2014] Kim, J., Cha, M., and Sandholm, T. (2014). Socroutes: safe routes based on tweet sentiments. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 179–182. ACM.

[Kung et al. 2014] Kung, K. S., Greco, K., Sobolevsky, S., and Ratti, C. (2014). Exploring universal patterns in human home-work commuting from mobile phone data. *PloS one*, 9(6):e96180.

[Li and Sun 2014] Li, C. and Sun, A. (2014). Fine-grained location extraction from tweets with temporal awareness. In *37th ACM SIGIR*, pages 43–52. ACM.

[Liu et al. 2011] Liu, X., Zhang, S., Wei, F., and Zhou, M. (2011). Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.

[Rettore et al. 2016] Rettore, P. H., Santos, B. P., Campolina, A. B., Villas, L. A., and Loureiro, A. A. (2016). Towards intra-vehicular sensor data fusion. *19th International Conference on ITS*.

[Ribeiro Jr et al. 2012] Ribeiro Jr, S. S., Davis Jr, C. A., Oliveira, D. R. R., and Meira Jr, W. (2012). Traffic observatory: a system to detect and locate traffic events and conditions using twitter. In *5th ACM SIGSPATIAL*.

[Yin and Du 2016] Yin, J. and Du, Z. (2016). Exploring multi-scale spatiotemporal twitter user mobility patterns with a visual-analytics approach. *ISPRS International Journal of Geo-Information*, 5(10):187.

[Zadeh 1984] Zadeh, L. A. (1984). Review of a mathematical theory of evidence. *AI Magazine*.

[Zheng et al. 2008] Zheng, Y., Li, Q., Chen, Y., Xie, X., and Ma, W.-Y. (2008). Understanding mobility based on gps data. In *10th ACM Ubiquitous computing*.