Detecção de Anomalias no Transporte Rodoviário Urbano*

Ana Beatriz Cruz[1], João Ferreira[1], Bernardo Monteiro[1], Rafaelli Coutinho [1], Fabio Porto[2], Eduardo Ogasawara[1]

[1] CEFET/RJ

[2]LNCC - DEXL Lab

anacruz@acm.org, joao.parana@acm.org

rafaelli.coutinho@cefet-rj.br, fporto@lncc.br, eogasawara@ieee.org

Abstract. *The constant increase in road traffic jams demands research related to urban mobility. These studies model the traffic analysis as a trajectory problem,* i.e.*, from the individualized analysis of vehicles that continuously send their geolocations. Such moving objects function as trajectory sensors and produce large amounts of data. It is possible to observe, however, an absence of studies related to spatial-temporal aggregations from trajectories of the urban vehicles. As a preliminary work on this subject, this paper establishes a baseline for anomaly identification in urban mobility, which may be found useful for developing new approaches that provide a better comprehension of urban mobility systems.*

Resumo. *O constante aumento de congestionamentos no tráfego rodoviário demanda pesquisas relacionadas a mobilidade urbana. Esses estudos modelam a análise do tráfego como um problema de trajetória,* i.e.*, a partir da análise individualizada de veículos que continuamente transmitem suas geolocalizações. Tais objetos móveis funcionam como sensores de trajetória e produzem grande quantidade de dados. Observa-se, entretanto, que há uma lacuna de estudos associados a agregações espaço-temporais das trajetórias dos veículos urbanos. Como um trabalho preliminar no assunto, este artigo estabelece uma base de comparação para identificação de anomalias em mobilidade urbana, que pode ser útil para o desenvolvimento de novas abordagens que provenham uma maior compreensão dos sistemas de mobilidade urbana.*

1. Introdução

O aumento da população mundial em áreas urbanas associado ao grande número de veículos presentes nas cidades provocam problemas como congestionamentos, acidentes e poluição [Ferreira et al., 2013; Chen et al., 2015]. Por conta disso, observou-se um aumento das pesquisas relacionadas a mobilidade urbana. Diversos estudos vêm sendo feitos motivados por compreender o comportamento do trânsito e minimizar o impacto de engarrafamentos.

Tais estudos têm por foco o comportamento das trajetórias de objetos móveis. Ferreira et al. [2013], por exemplo, usaram dados emitidos por GPS de táxis para criar uma

*Os autores agradecem a CAPES, CNPq e FAPERJ pelo financiamento parcial do projeto.

ferramenta que favorece a análise visual da mobilidade da cidade de Nova York. Verhein and Chawla [2008] usaram dados emitidos por celulares, RFID e rastreamento por satélite para identificar trajetórias de trânsito intenso a partir da movimentação dos veículos. Pan et al. [2013] também estudaram as variações nas movimentações das trajetórias para observar comportamentos anômalos em algumas rotas.

Objetivando análises de comportamento mais sistêmicas, Andrienko and Andrienko [2008] propuseram métodos de agregação espaço-temporal, consolidando os dados de trajetória em mapas que representam principais fluxos no espaço urbano. Esses mapas de fluxo podem ser descritos por séries espaço-temporais das regiões (objetos permanentes) que ajudam a entender comportamentos da mobilidade urbana pela perspectiva das regiões [Adrienko and Adrienko, 2011].

Neste contexto, observa-se uma lacuna de estudos pela ótica das séries espaço-temporais de objetos permanentes que possam trazer uma melhor compreensão ou uma visão complementar do tráfego. Como um trabalho preliminar no assunto, este artigo estabelece uma metodologia para identificação de anomalias em dados agregados de mobilidade urbana. Tal abordagem estabelece uma base de comparação que pode ser útil para o desenvolvimento de trabalhos futuros que objetivam a sistematização e análise mais aprofundada dessas anomalias.

Além desta introdução, o trabalho está organizado em mais quatro seções. Na seção 2, são apresentados conceitos gerais necessários para o entendimento do problema e da metodologia aplicada. Em seguida, na seção 3, apresenta-se a metodologia adotada nesse trabalho. A seção 4 descreve uma avaliação preliminar da proposta. A seção 5 finaliza com as considerações finais.

2. Formalização

As séries espaço-temporais são definidas como sequências de observações de objetos que contêm dados sobre o local e momento em que as coletas foram realizadas [Cressie and Wikle, 2015]. Os objetos com localização fixa são classificados como permanentes (interpretados como sensores fixos), enquanto aqueles cujas localizações variam com o tempo são classificados como móveis (interpretados como trajetórias). A trajetória é o modelo de dados mais aplicado a problemas relacionados ao tráfego [Chen et al., 2015]. Entretanto, estudos que se relacionam mais diretamente com o tema deste trabalho são aqueles em que se agregam informações do objeto [Tao et al., 2004], gerando séries espaço-temporais associadas a objetos permanentes.

Uma trajetória de um objeto tr_k é descrita como sequência = $<\hat{tr}_1, \cdots, \hat{tr}_{co}>$ de pares $\hat{tr}_o = (p, v)$, onde $\hat{tr}_o.v$ corresponde a um valor ($\hat{tr}_o.v \in \mathbb{R}$) e $\hat{tr}_o.p$ a uma posição ($\hat{tr}_o.p \in \mathbb{R}^2$) no tempo o. O número de observações (co) em uma trajetória é dado por $|tr_k|$. Define-se também um *dataset* de trajetórias $TR = \cup_{k=1,ck} tr_k$, representado pelo conjunto de trajetórias tr_k em um sistema ao longo de um período de análise. O número de trajetórias (ck) em TR é dado por $|TR|$.

Durante o processo de agregação, é necessário estabelecer partições espaciais e temporais. Considere $S = \cup_{i=1,ci} \hat{s}_i$, regiões bi-dimensionais $|S| = ci, \hat{s}_i = [(xl_i, yl_i), (xu_i, yu_i)]$ que particionam o espaço S em ci regiões retangulares definidas pelos limites inferiores xl_i, yl_i e superiores xu_i, yu_i para os eixos x e y. Considere também

$T = \cup_{j=1,cj}\ \hat{t}_j, |T| = cj, \hat{t}_j = [(j-1)\cdot ma + 1, j \cdot ma)]$ um particionamento do tempo T em cj intervalos de duração ma.

Seja $ST = \cup_{i=1,ci}\ st_i, st_i = <\hat{u}_{i,1}, \cdots, \hat{u}_{i,cj}>$ um conjunto de séries espaço-temporais st_i provenientes das ci partições espaciais. Cada st_i contém uma série temporal de objeto permanente de tamanho cj provenientes das cj partições no tempo. Cada uma das observações $\hat{u}_{i,j}$ é produzida a partir da agregação espaço-temporal das trajetórias dos objetos. A Equação 1 descreve a agregação, aplicando-se a função de agregação θ sobre os valores $tr_k.\hat{tr}_o.v$ dos objetos que estavam presentes no particionamento espaço-temporal $\hat{s}_i, \hat{t}_j$.

$$\forall \hat{s}_i \in S, \hat{t}_j \in T\ (\hat{u}_{i,j} = \ _{\hat{s}_i,\hat{t}_j}\Gamma_{\theta(tr_k.\hat{tr}_o.v)}(\sigma_{o\in\hat{t}_j, tr_k.\hat{tr}_o.p\in\hat{s}_i} tr_k.\hat{tr}_o)) \quad (1)$$

Seja st_i uma série espaço-temporal de ST. Considere ma o intervalo de particionamento do tempo em minutos, tem-se que um dia contém $obs = 1440/ma$ observações. Por simplicidade, considere que ma é divisor de 1440. Para $w \in [1..obs]$, as observações de $\hat{u}_{i,w+d\cdot ma}$, para $d \in N$ representando dias, correspondentes às observações realizadas no mesmo horário nos diferentes dias. Fazendo-se variar w e tendo-se d um conjunto de dias, tem-se as $st_{i,w,d}$ as distribuições nos diferentes horários (w) caracterizados por cada séries st_i em d.

A partir destas distribuições, pode-se estabelecer os valores esperados $\overline{st_{i,w,d}}$ e intervalos típicos para as observações na região i, no horário w ao longo dos dias d. Assumindo-se distribuição não normal, o intervalo típico é descrito por $IT_{i,w,d} = [q_1 - 3\cdot IQR, q_3 + 3\cdot IQR]$ [Larsen and Marx, 2005], onde q_1 é o primeiro quartil, q_3 é o terceiro quartil e IQR é a distância interquartil da distribuição $st_{i,w,d}$. Os valores acima ou abaixo destes intervalos podem ser estatisticamente interpretados como anomalias.

3. Metodologia

O presente trabalho objetiva prover um método básico para identificar anomalias no sistema de transporte urbano a partir das séries espaço-temporais derivadas de agregações das geolocalizações emitidas minuto a minuto pelos ônibus do Rio de Janeiro. Os ônibus atuam, portanto, como sensores de trajetória. Os seus dados podem ser obtidos no Portal de Dados Abertos da Prefeitura do Rio de Janeiro [DataRio, 2016]. O processo para identificação dessas anomalias é dividido em cinco atividades principais: (i) seleção de objetos permanentes, (ii) agregação espaço-temporal, (iii) definição de intervalo típico, (iv) identificação das anomalias e (v) análise dos resultados.

A seleção dos objetos permanentes (regiões) é caracterizada pela divisão do município do Rio de Janeiro em um reticulado de 500 por 500 metros projetado sobre a área estudada. Devido as particularidades da geografia de algumas cidades, como o Rio de Janeiro, uma seleção foi feita a fim de descartar quadrantes localizados em regiões em que não há tráfego de ônibus. A etapa de agregação espaço-temporal foi realizada considerando este reticulado e o particionamento temporal de dez minutos. A agregação espaço-temporal considera a velocidade média, quantidade e lista de ônibus e de linhas presentes na região para cada intervalo de dez minutos. Ao fim dessa etapa, são obtidas séries espaço-temporais para cada região. Posteriormente, é feito o cálculo do intervalo típico de cada região que serve de base de comparação para as observações realizadas. Os

valores de velocidade identificados fora desse intervalo indicam anomalias nos padrões do trânsito.

Este processo foi implementado como um workflow. Dado o cenário de larga escala de dados e a necessidade de processamento de alto desempenho, o workflow foi desenvolvido em Apache *Spark*. O encadeamento do workflow foi implementado em Scala. As atividades foram escritas em Python e R [Ferreira et al., 2017] e invocadas a partir da especificação em Scala.

4. Avaliação Experimental

A Metodologia descrita na seção 3 foi aplicada sobre o período de dois meses: julho e agosto de 2014. O mês de julho teve em média 2.961.726 observações diárias, enquanto que a média de observações diárias de agosto é de 3.403.180. Inicialmente, para cada mês, foram calculados os intervalos típicos. A metodologia de identificação de anomalias aplicada sobre o mês de julho resultou na identificação de 42.050 anomalias. Esse volume representa aproximadamente 1,03% das observações. Em agosto, foram identificadas 38.843 anomalias, volume que representa aproximadamente 1,07% da base comparada. Também aplicamos os intervalos típicos computados em julho para servir de base de comparação relativa com as observações do mês de agosto. Desta forma, foram identificadas 64.041 anomalias, correspondente a 1,76% do volume da base comparada.

A Figura 1 ilustra a quantidade de anomalias por área de planejamento da cidade nos períodos estudados. Embora as análises sobre meses de julho e agosto individualmente indiquem a maior concentração de anomalias na área de planejamento da Zona Norte do Rio de Janeiro, a comparação relativa entre agosto e julho mostrou que o maior número de anomalias ocorreu na área de planejamento da Zona Oeste, seguida pela Zona Norte. Na Zona Norte, o maior número de anomalias aponta para um aumento da velocidade média, enquanto que na Zona Oeste, a maioria das anomalias identificadas indicam o aumento de engarrafamentos. As anomalias que indicam redução da velocidade na Zona Oeste, concentram-se principalmente nos bairros de Guaratiba, Santa Cruz e Campo Grande, que, juntos, representam aproximadamente 71,93% das anomalias de lentidão na Zona Oeste e 23,8% das anomalias de lentidão de todo o Rio de Janeiro.

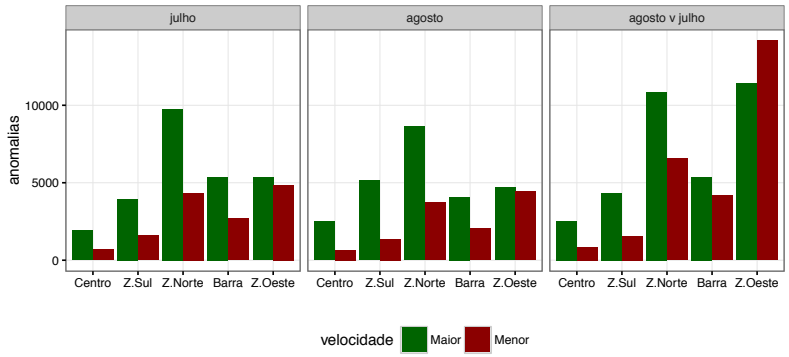

Figura 1. Anomalias identificadas por área de planejamento

As anomalias abertas por dia da semana podem ser vistas na Figura 2. Um resultado esperado é que ocorram aumentos significativos de velocidades aos domingos. No mês de julho, em especial, anomalias que indicam o aumento da velocidade média são

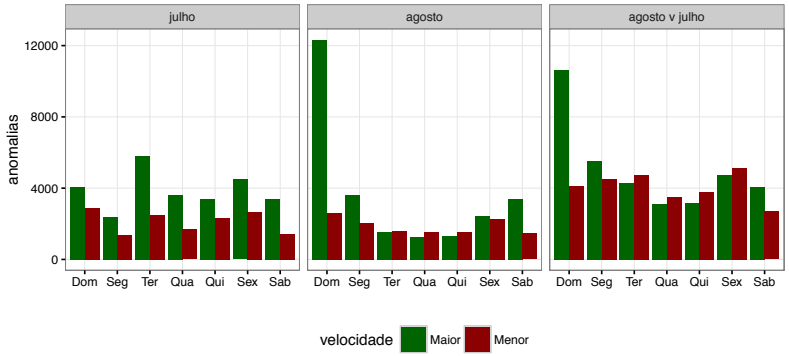

Figura 2. Anomalias identificadas por dia da semana

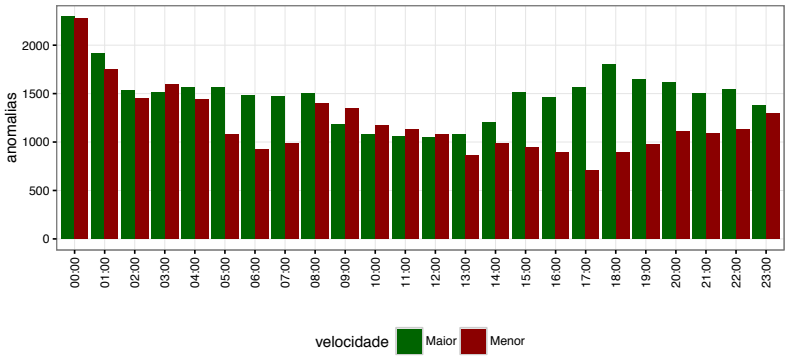

Figura 3. Anomalias identificadas por faixa de horário (ago v julho)

maiores nas terças. Esse comportamento específico do mês de julho ocorreu por causa da Copa do Mundo, que teve alguns de seus jogos no Rio de Janeiro. Nesse mês, os dias 08/07/2014 e 04/07/2014 foram os que apresentaram mais anomalias de maior velocidade. O dia 08/07/2014 (terça feira) foi o dia do jogo entre Brasil e Alemanha nas semifinais da Copa do Mundo. O dia 04/07/2014 (sexta feira) foi feriado devido à Copa do Mundo. Foi identificado ainda um volume de anomalias associados à diminuição de velocidade no domingo superior ao observado no mês de agosto. Esse comportamento anômalo foi decorrente dos protestos que ocorreram na cidade no dia 06 de julho, da 14ª parada gay que ocorreu no dia 20 de julho e da demolição do último segmento do Elevado da Perimetral, interdição da Avenida Rodrigues Alves e alteração no percurso de diversas linhas de ônibus no dia 27 de julho.

As anomalias abertas por faixa de horário, comparando-se agosto versus julho, podem ser vistas na Figura 3. A partir desta comparação, foram identificadas anomalias que indicam o aumento da velocidade média em horários considerados de pico como o intervalo entre cinco horas da tarde e oito horas da noite. No período entre seis e oito horas da manhã, apesar do número de anomalias que indicam velocidades maiores ser superior ao número de anomalias que indicam diminuição da velocidade média, pode-se observar o aumento gradual do número de anomalias que indicam velocidades mais lentas. Esse aumento gradual de anomalias culmina na maior proporção de anomalias de diminuição de velocidade a partir das nove horas da manhã, normalizando após meio dia. Esse comportamento pode ser visto no gráfico ilustrado na Figura 3 e reflete o comportamento de formação de engarrafamentos no horário de pico da manhã.

5. Considerações finais

O presente trabalho apresenta um método básico para identificação de anomalias no comportamento do trânsito rodoviário analisados por meio de agregações espaço-temporais. Tomou-se como base os dados abertos do município do Rio de Janeiro. A metodologia proposta foi implementada em Apache Spark e identificou algo equivalente a 1% de anomalias por mês estudado. Tal abordagem estabelece uma base de comparação que pode ser útil para o desenvolvimento de trabalhos que objetivam a sistematização e análise mais aprofundada dessas anomalias. Por exemplo, observou-se que há diferentes tipos de anomalias. Alguns têm comportamentos pontuais oriundos de fenômenos singulares (como o dia do jogo do Brasil versus Alemanha na semi-final da Copa do Mundo de 2014), enquanto outros são recorrentes e descrevem comportamentos mais sistêmicos. Neste caso, abre-se espaço para explorar técnicas de padrões frequentes capazes de aglutinar as anomalias recorrentes como também aquelas que são perenes em intervalos de tempo e espaço contíguos.

Referências

Adrienko, N. and Adrienko, G. (2011). Spatial generalization and aggregation of massive movement data. *IEEE Transactions on visualization and computer graphics*, 17(2):205–219.

Andrienko, G. and Andrienko, N. (2008). Spatio-temporal aggregation for visual analysis of movements. In *Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on*, pages 51–58. IEEE.

Chen, W., Guo, F., and Wang, F.-Y. (2015). A survey of traffic data visualization. *Intelligent Transportation Systems, IEEE Transactions on*, 16(6):2970–2984.

Cressie, N. and Wikle, C. K. (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

DataRio (2016). Portal de dados abertos da prefeitura do Rio de Janeiro. Technical report, `http://data.rio/`.

Ferreira, J., Gaspar, D., Monteiro, B., Silva, A. B., Porto, F., and Ogasawara, E. (2017). Uma Proposta de Implementação de Álgebra de Workflows em Apache Spark no Apoio a Processos de Análise de Dados. In *Brazilian e-Science Workshop*.

Ferreira, N., Poco, J., Vo, H. T., Freire, J., and Silva, C. T. (2013). Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2149–2158.

Larsen, R. J. and Marx, M. L. (2005). *An Introduction to Mathematical Statistics and Its Applications*. Prentice Hall, Upper Saddle River, N.J, 4 edition edition.

Pan, B., Zheng, Y., Wilkie, D., and Shahabi, C. (2013). Crowd Sensing of Traffic Anomalies Based on Human Mobility and Social Media. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL'13, pages 344–353, New York, NY, USA. ACM.

Tao, Y., Kollios, G., Considine, J., Li, F., and Papadias, D. (2004). Spatio-temporal aggregation using sketches. In *Data Engineering, 2004. Proceedings. 20th International Conference on*, pages 214–225.

Verhein, F. and Chawla, S. (2008). Mining spatio-temporal patterns in object mobility databases. *Data mining and knowledge discovery*, 16(1):5–38.