Detecção de Anomalias Frequentes no Transporte Rodoviário Urbano*

Ana Beatriz Cruz[1], João Ferreira[1], Diego Carvalho[1], Eduardo Mendes[4] Esther Pacitti[3], Rafaelli Coutinho[1], Fabio Porto[2], Eduardo Ogasawara[1]

[1] CEFET/RJ

[2]LNCC - DEXL Lab

[3]FGV

[4]Inria & University of Montpellier

anacruz@acm.org, joao.parana@acm.org, d.carvalho@ieee.org,

eduardo.mendes@fgv.br, renato.souza@fgv.br, Esther.Pacitti@inria.fr,

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

Abstract. *The growth of urban population and, consequently, the number of vehicles causes the increase of traffic jams and emission of polluting gases. In this context, we observe the intensification of papers that aim to identify bottlenecks and their causes. These papers propose methodologies that use trajectory data model and aim to explain systemic behaviors. This article proposes the identification and classification of anomalies in the urban road transport system from space-time aggregations to permanent objects. The methodology consists of pre-processing of data, identification of anomalies, identification, and classification of frequent patterns. Through it, we can identify the systemic and specific behaviors on the urban transit of Rio de Janeiro.*

Resumo. *O crescimento da população urbana e, consequentemente, do número de veículos provoca o aumento de engarrafamentos e da emissão de gases poluentes. Nesse contexto, observa-se a intensificação de pesquisas que buscam identificar engarrafamentos e suas causas. Estas pesquisas propõem metodologias que usam modelo de dados de trajetória e visam explicar comportamentos sistêmicos. Este artigo propõe a identificação e a classificação de anomalias no sistema de transporte rodoviário urbano a partir de agregações espaço-temporais a objetos permanentes. A metodologia consiste do pré-processamento dos dados, identificação de anomalias, identificação e classificação de padrões frequentes. Por meio dela, é possível identificar comportamentos sistêmicos e pontuais do trânsito urbano do Rio de Janeiro.*

1. Introdução

Em 2007, pela primeira vez existiam mais pessoas vivendo em áreas urbanas do que em zonas rurais, resultado de uma urbanização expressiva que se impulsionou desde a década

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

de 1950 [United Nations, 2014]. Atualmente, mais da metade da população mundial vive em áreas urbanas exigindo um replanejamento dos serviços públicos das zonas urbanas para provê-los de maneira sustentável e duradoura. Outrossim, os desafios relacionados ao desenvolvimento eficiente e sustentável dos serviços de transporte passaram a ser investigados [Chen et al., 2015], levando à uma intensificação das pesquisas que conjugam a análise de dados de transporte à mobilidade urbana, com o objetivo de se identificar fenômenos causadores dos estrangulamentos do transporte, como os engarrafamentos.

Os estudos sobre os estrangulamentos fazem análise de dados coletados principalmente a partir de sistemas GP/dispositivos móveis embarcados em veículos que participam do fluxo, como táxis [Ferreira et al., 2013] e ônibus [Bierlaire et al., 2013]. Os dados coletados são frequentemente modelados como trajetórias de objetos móveis, pontos de início ou fim de movimento. Para análises mais sistêmicas, métodos de agregação espaço-temporal são usados para associar as observações às posições geográficas predefinidas (objetos permanentes) e esse tipo de agregação reduz significativamente o volume de dados de trajetórias [Tao et al., 2004].

Neste contexto, observa-se a necessidade de um estudo mais aprofundado sobre ótica das séries espaço-temporais de objetos permanentes que possam trazer uma melhor compreensão do tráfego [Cruz et al., 2017]. Este trabalho tem por objetivo identificar e classificar anomalias em dados agregados de mobilidade urbana. Para isso, uma técnica de identificação de anomalias nos comportamentos do trânsito é aplicada sobre os dados agregados por regiões predefinidas. Para extrair conhecimentos destas anomalias, um novo método de classificação de padrões frequentes é proposto, permitindo identificar padrões anômalos inesperados e esperados.

Além dessa introdução, o trabalho se divide em quatro outras seções. A seção 2 apresenta conceitos essenciais para o entendimento do problema e da metodologia adotada. Na seção 3, apresenta-se a metodologia aplicada. A seção 4 descreve uma avaliação da proposta. Finalmente, a seção 5 apresenta as conclusões e os próximos passos.

2. Fundamentação Teórica

As séries espaço-temporais são definidas como sequências de observações de objetos que contêm dados sobre o local e momento das coletas [Cressie and Wikle, 2015]. As observações podem ser emitidas por objetos permanentes ou móveis. Os objetos permanentes possuem localização fixa (sensores fixos) e os objetos móveis apresentam localizações que variam com o tempo (trajetória). O modelo de dados espaço-temporal mais aplicado a problemas relacionados ao tráfego é o de trajetória [Chen et al., 2015]. Os dados emitidos por sensores de posicionamento, como o GPS, possuem informações de latitude, de longitude e do momento da coleta. Desta forma, a sequência de dados coletados por sensores de posicionamento configura naturalmente uma trajetória, sem a necessidade de pré-processamento. Entretanto, estudos que se relacionam mais diretamente com o tema deste trabalho são aqueles em que agregam informações do objeto [Tao et al., 2004], gerando séries espaço-temporais associadas a objetos permanentes.

A partir de observações sobre um sistema é possível encontrar características específicas. A recorrência dessas características torna-as esperadas. Desvios significativos nas propriedades das características esperadas do sistema são consideradas anomalias [Aggarwal, 2016]. No contexto de mobilidade urbana, anomalias podem indicar

mudanças no comportamento por engarrafamentos ou aumento da velocidade. Elas podem ser causadas por acidentes, eventos, obras, operações de controle como *blitz* e Lei Seca, protestos, desastres e feriados. São difíceis de serem encontradas e interpretadas devido ao grande volume de dados e ao grande número de ruídos que compõem o conjunto de dados analisado [Lakhina et al., 2004]. Dessa forma, a identificação de anomalias é frequentemente feita sobre dados de trajetória.

A mineração de padrões frequentes é um dos métodos para identificar padrões que ocorrem com frequência no conjunto de dados analisado Han et al. [2011]. Esses padrões podem ser itens, sequências ou estruturas. Em todos os casos, por meio da mineração de padrões frequentes são identificadas regras de associação e correlações. Uma regra de associação é uma implicação de formato $X \rightarrow Y$. Seja $I = \{i_1, i_2, ..., i_n\}$ um conjunto de todos os itens, o antecedente da implicação (X) e o consequente (Y) são conjuntos de itens em I no qual nenhum item em X pertencente a Y e vice-versa. Logo, $X \cap Y = \emptyset$. A partir das regras de associação são identificados itens que ocorrem com frequência em uma mesma transação. Para que regras de associação sejam consideradas importantes, condições para suporte, confiança e correlação *lift* geralmente devem ser satisfeitas.

Entre as técnicas de mineração de padrões frequentes mais difundidas, destaca-se o algoritmo Apriori [Han et al., 2011]. Ele se baseia no princípio de que um conjunto de itens será frequente se todos os seus subconjuntos também forem. Em mobilidade urbana, algoritmos de identificação de padrões sequenciais frequentes são geralmente aplicados como base para identificar padrões frequentes em trajetórias [Giannotti et al., 2007].

Diversas medidas de classificação de padrões frequentes já foram propostas e elas se dividem em objetivas e subjetivas. As técnicas objetivas baseiam-se em propriedades estatísticas e as técnicas subjetivas, em conhecimentos de especialistas sobre o domínio [Mcgarry, 2005]. Os valores de suporte, confiança e *lift* são obtidos a partir de cálculos estatísticos e são classificadas como objetivas. As técnicas objetivas frequentemente retornam algumas regras que já são conhecidas ou triviais. Por outro lado, apesar da eficácia de técnicas subjetivas e sua influência na qualidade da classificação de padrões, a técnica é custosa devido à dificuldade e à complexidade para aquisição de conhecimentos prévios. Nesse contexto, pesquisas tem sido desenvolvidas aplicando-se uma combinação de técnicas subjetivas e objetivas a fim de extrair as vantagens de cada uma.

3. Metodologia

Este trabalho tem como objetivo identificar e classificar anomalias que ocorrem no sistema de transporte rodoviário urbano a partir das séries espaço-temporais derivadas de agregações das geolocalizações emitidas pelos ônibus da cidade do Rio de Janeiro[1]. Para isso, o processo é dividido em cinco etapas: pré-processamento (1) *Extract, Transform and Load* (ETL) e (2) agregação espaço-temporal; (3) identificação de anomalias; (4) identificação de padrões frequentes e (5) avaliação de padrões frequentes.

No pré-processamento (etapa 1), os dados são tratados até que possam ser minerados. Duas etapas principais o compõem: ETL e agregação espaço-temporal. A etapa ETL é responsável pela extração e limpeza dos dados abertos de mobilidade da cidade do Rio de Janeiro. Por meio da agregação espaço-temporal (etapa 2), os dados são convertidos

[1]Os dados são obtidos pelo Portal de Dados Abertos da Prefeitura do Rio de Janeiro: http://data.rio/

em séries espaço-temporais associadas a objetos permanentes, resultando em uma visão alternativa dos dados de mobilidade. A etapa seguinte (etapa 3) consiste na identificação de anomalias para compreensão de impactos de eventos, mudanças no tráfego e feriados. Neste presente trabalho, as anomalias nas séries espaço-temporais são identificadas com a aplicação de conceitos estatísticos conforme proposto em Cruz et al. [2017].

De modo a extrair conhecimento útil, as anomalias são estudadas por meio da aplicação da técnica de identificação de padrões frequentes. Os padrões identificados, chamados de regras de associação, configuram uma implicação lógica, na qual o antecedente é chamado condição e o consequente é chamado consequência. O algoritmo de identificação de padrões frequentes adotado (etapa 4) é o Apriori. O algoritmo de identificação de padrões frequentes é executado sobre as anomalias do transporte rodoviário variando-se os valores de suporte e confiança. Para cada valor de cada parâmetro selecionado, são avaliados os números de regras de associação resultantes e a qualidade dessas regras. A identificação de padrões frequentes é aplicada a duas granularidades diferentes: anomalias identificadas ao longo do ano e em cada mês individualmente.

Para classificar os padrões frequentes identificados, um método adaptado da técnica proposta por Liu et al. [2000], que combina abordagens subjetivas e objetivas de classificação, é proposto (etapa 5). Em Liu et al. [2000], a seleção de regras esperadas é feita por meio da informação dos conhecimentos de especialistas. De modo a não depender dos conhecimentos de um especialista, a primeira etapa da técnica de classificação de padrões proposta consiste em encontrar padrões que ocorram em uma visão anual das anomalias. Dessa forma, aplica-se o algoritmo de identificação de padrões frequentes sobre as anomalias identificadas ao longo de um ano. As regras de associação produzidas na visão anual são consideradas regras esperadas se possuem apenas um termo na condição e se o suporte e o *lift* são superiores aos valores mínimos passados como parâmetro.

A classificação de padrões frequentes é, então, aplicada a menores granularidades (em cada mês) segundo os valores de *lift* e de regras não esperadas (RNE), que mensura o quão não esperada é uma regra. Para calcular o valor de RNE, as condições (antecedentes) e consequências das regras são avaliadas, recebendo valor entre 0 e 1 para sinalizar se o antecedente (AE) e a consequência (CE) são esperados. Para isso, os termos que compõem as condições e consequências das regras produzidas por meio da identificação de padrões frequentes de menor granularidade são comparadas com os termos das regras esperadas produzidas na visão anual para gerar os valores AE e CE. O valor de mensuração de regras esperadas (RE) é resultante da média aritmética dos valores AE e CE. O valor de RNE é o inverso do valor de RE. Assim, quanto maiores são os valores de RNE e do *lift*, maior será a relevância da regra.

4. Avaliação Experimental

Essa seção descreve os resultados obtidos pela abordagem proposta sobre dados coletados de maio de 2014 a abril de 2015. No ano de 2014, observaram-se medidas que visavam melhorias na dinâmica do trânsito no Rio de Janeiro e o Brasil sediou a Copa do Mundo. A metodologia de identificação de anomalias aplicada sobre período avaliado resultou na identificação de 398.662 anomalias. Com exceções de julho e de agosto de 2014, o número de anomalias identificadas em todos os meses é inferior à 1% do volume de dados total de cada mês. De modo geral, domingos possuem mais anomalias

de velocidade maior que o intervalo típico em todos os meses. O mês de maio de 2014 teve maior número de anomalias que indicam velocidades inferiores ao intervalo típico. Nesse período, ocorreram diversas manifestações, greves de rodoviários e muitas obras na cidade do Rio de Janeiro ainda não haviam sido finalizadas.

A técnica de alisamento por média foi aplicada nos dados de horário e velocidade para discretizá-los e cada unidade espacial foi associada ao bairro no qual está localizado. Foram usados o suporte mínimo de 1% e a confiança mínima de 39% no algoritmo *Apriori* para identificação de padrões. O *Apriori* foi executado para anomalias identificadas ao longo do ano e em cada mês individualmente. Foram produzidas 292 e 2614 regras de associação para as perspectivas anual e mensal, respectivamente.

Em problemas de padrões frequentes, espera-se encontrar regras inesperadas, novas ou que tenham utilidade em tomadas de decisões e ações de especialistas. Algumas regras de associação produzidas não agregam conhecimentos relevantes. Por isso, aplica-se a metodologia de classificação das regras. As regras esperadas apresentaram *lift* > 1 e suporte relativamente alto. Por esse motivo, regras com apenas um termo no antecedente, *lift* ≥ 1 e suporte $> 0,05$ foram classificadas como regras esperadas. Foram identificadas 23 regras esperadas. Por exemplo, a regra $\{velocidade = Lenta\} => \{tipo = Menor\}$ é esperada. Ela indica que quando a velocidade é *Lenta*, a anomalia é do tipo *Menor*, ou seja, inferior ao intervalo típico. Elas servem de base para classificar as regras produzidas com a aplicação do *Apriori* sobre anomalias do ano e de cada mês individualmente.

A Figura 1 ilustra o número de regras por classificação sob perspectiva de mês e ano. A análise sobre as regras de associação para anomalias identificadas ao longo do ano identificou 70 regras como inesperadas. Aproximadamente 70% das regras são consideradas esperadas. A mesma análise foi feita para anomalias identificadas a cada mês individualmente. Para esta perspectiva, apenas 583 regras foram consideradas completamente inesperadas, ou seja, $RNE = 1$, e 1.282 regras, onde $0 <$RNE< 1.

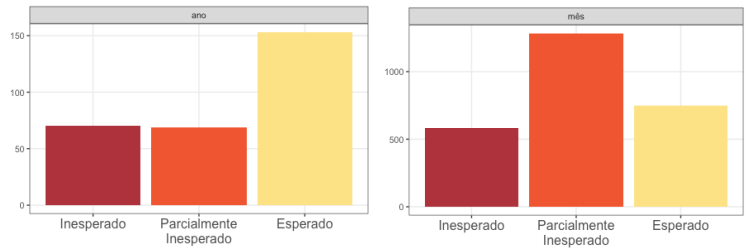

Figura 1. Número de regras por classificação sob perspectiva de mês e ano.

Para fazer uma análise das regras, elas foram ordenadas de forma crescente pelo valor de RNE e número de termos na condição, e de forma decrescente de acordo com o $lift$. Com essa ordenação, espera-se que regras inesperadas e mais generalistas estejam no topo. Seguindo esse critério, as regras classificadas como mais relevantes tem informações de datas ou de bairros. Em relação aos bairros, destacam-se Vila Militar, Jardim Sulacap, São Cristóvão, Manguinhos e Guaratiba com grande número de anomalias que indicaram velocidade inferior ao intervalo típico. Vila Valqueire, por sua vez, possui anomalias que indicaram velocidade superior ao intervalo típico, com 92% de confiança.

Por meio da análise das regras que contém informação de data na condição, a

regra classificada como mais relevante é o dia 29/06/2014 (domingo), que tem como consequência a velocidade inferior ao intervalo típico. A regra, além de ir contra o tipo de anomalia esperada para um domingo, possui $lift = 2,17$ e 91% de confiança. Uma análise sobre reportagens referentes a data em questão indicou que, além de estar ocorrendo a Copa do Mundo, houve um protesto na zona sul do Rio de Janeiro e uma operação conjunta da Secretaria de Ordem Pública e Secretaria de Municipal de Transportes de fiscalização de táxis na Rodoviária Novo Rio.

5. Conclusão

Este trabalho propõe uma metodologia para identificar e classificar as anomalias no comportamento do trânsito analisadas por agregações espaço-temporais. Foram usados dados da cidade do Rio de Janeiro. A metodologia proposta identificou características das principais anomalias e classificou-as como esperadas ou inesperadas. A aplicação da identificação de regras esperadas se mostrou eficiente não apenas para classificação das anomalias pontuais, como para entendimento de comportamentos sistêmicos do trânsito. Existem ainda, oportunidades para trabalhos futuros com a aplicação da metodologia proposta em outros conjuntos de dados, como dados de sistemas de transporte privado e táxis. Diferentes técnicas de identificação de anomalias também podem ser exploradas a fim entender como elas podem impactar nos resultados e nas análises.

Referências

Aggarwal, C. C. (2016). *Outlier Analysis*. Springer, New York, NY, 2nd edition.

Bierlaire, M., Chen, J., and Newman, J. (2013). A probabilistic map matching method for smartphone GPS data. *Transportation Research Part C: Emerging Technologies*, 26:78–98.

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.

Cruz, A. B., Ferreira, J., Monteiro, B., Coutinho, R., Porto, F., and Ogasawara, E. (2017). Detecção de anomalias no transporte rodoviário urbano. In *Proceedings of the 32nd Brazilian Symposium on Databases (SBBD)*, pages 240–245.

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.

Giannotti, F., Nanni, M., Pinelli, F., and Pedreschi, D. (2007). Trajectory pattern mining. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 330–339.

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, Haryana, India; Burlington, MA, 3 edition.

Lakhina, A., Crovella, M., and Diot, C. (2004). Diagnosing network-wide traffic anomalies. In *ACM SIGCOMM Computer Communication Review*, pages 219–230. ACM.

Liu, B., Hsu, W., Chen, S., and Ma, Y. (2000). Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems and their Applications*, 15(5):47–55.

Mcgarry, K. (2005). A survey of interestingness measures for knowledge discovery. *The Knowledge Engineering Review*, 20(1):39–61.

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

United Nations (2014). World urbanization prospects. https://www.un-ilibrary.org/content/publication/527e5125-en.