O QUE É RELEVANTE NA PREDIÇÃO DAS CONDIÇÕES DO TRÁFEGO A PARTIR DE DADOS DE TWITTER E OUTRAS FONTES HETEROGÊNEAS? UM ESTUDO PRELIMINAR SOBRE A CIDADE DE PORTO ALEGRE

Rhuam Sena Estevam
Jorge C. Chamby-Diaz
Ana L. C. Bazzan
Universidade Federal do Rio Grande do Sul
Instituto de Informática

RESUMO
Várias pesquisas sugerem que há relação entre observações feitas na malha viária (e.g., nível de congestionamento) e informações que podem ser extraídas da Internet. Porém, não é claro que tipo de informação é relevante. Este trabalho reporta os resultados de um estudo preliminar usando seleção de atributos. Foi utilizado o ganho de informação sobre bases de dados extraídas da Internet, como contas no Twitter. Com os dados aqui considerados, é possível concluir que atributos referentes a condições meteorológicas apresentam maior importância, seguidos por localidade.

ABSTRACT
Several works suggest that there is a relationship between observations about the traffic network (e.g., congestion level) and information that can be extracted from the Internet. However, it is not clear which information is indeed relevant. The present work aims at investigating this using feature selection. Gain of information was used over datasets extracted from the Internet, such as Twitter accounts. In this preliminary study, we concluded that features related to weather have higher importance, followed by location.

1. INTRODUÇÃO

Prever o comportamento do tráfego, bem como incidentes de várias naturezas na rede viária pode ser útil no seu planejamento e operação. Embora alguns trabalhos citem diversas possibilidades de se realizar tal agenda, a grande maioria considera apenas informação sobre a própria rede viária como entrada para os métodos propostos. Porém, alguns autores Wang *et al.* (2014); Yazici *et al.* (2017) sugerem a obtenção de informações relevantes a partir da Internet, incluindo aí redes sociais. Segundo Pereira *et al.* (2014), qualquer informação semântica que possa ser associada com as observações feitas no sistema de sensoriamento de tráfego – como câmeras, sondas GPS, laços induzidos – pode ser denominada de **contexto**. Pereira *et al.* (2014) descrevem ainda a importância da aplicação dessas informações de contextos – como condições meteorológicas, eventos esportivos, culturais e outros, obras, etc. – para explicar e prever fenômenos relacionados ao trânsito.

Os dados relativos aos diversos contextos são formados por vários **atributos**. Por exemplo, o contexto relacionado às condições meteorológicas envolve atributos como precipitação (tipo e volume), temperatura, etc. Entretanto, para se realizar uma boa predição, seja através de métodos de classificação ou outros métodos, é importante se realizar, antes, um pré processamento que envolva o estudo sobre a qualidade de tais atributos. Isto se deve ao fato de que alguns atributos podem ser redundantes ou até mesmo irrelevantes e/ou causar *overfitting*. Desta forma, tal pré processamento visa não apenas melhorar a acurácia da classificação (e, portanto, da predição), mas também o desempenho computacional e, não menos importante, melhorar a compreensibilidade do modelo final a ser apresentado ao usuário. Em suma, é recomendável investigar como e quais atributos dos diversos contextos considerados podem estar relacionados a determinados comportamentos observados na malha viária, antes de utilizar qualquer técnica

de predição.

O restante deste texto está organizado da seguinte forma. A Seção 2 discute trabalhos anteriores, os desafios, bem como descreve brevemente conceitos relacionados a aprendizado de máquina, classificação e seleção de atributos. A Seção 3 descreve os materiais e métodos empregados. A Seção 4 descreve os experimentos realizados e os resultados observados. A última seção traz considerações finais e possíveis extensões.

2. TRABALHOS RELACIONADOS E REFERENCIAL TEÓRICO

Conforme detalhado a seguir, diversos trabalhos apontam condições meteorológicas como sendo um fator responsável por alteração no tráfego e ocorrência de acidentes. Porém, a maioria desses trabalhos se concentra em relações parciais, ou seja, considera apenas a relação das condições meteorológicas com os acidentes ou com o congestionamento. Koetse e Rietveld (2009) discutem uma série de trabalhos que explicam as relações entre condições meteorológicas e o efeito no transporte. Em dias de neve e forte chuva, as relações fluxo-ocupação e velocidade-fluxo são fortemente alteradas e há um grande impacto na operação (Ibrahim e Hall, 1994). Maze *et al.* (2006) citam a influência do vento e da temperatura na variação de velocidade média, porém os autores consideram essas informações menos dignas de nota do que a influência de uma chuva forte ou nevasca. A ocorrência de acidentes também é objeto de estudo de alguns autores. Neve e vento aparecem no trabalho de Edwards (1996) como fatores que influenciam o aumento de acidentes. A chuva é uma fator que por um lado, reduz a gravidade dos acidentes e, por outro, aumenta a sua frequência (Brodsky e Hakkert, 1988; Eisenberg, 2004; Qiu e Nixon, 2008).

Outros contextos que explicam a condição do tráfego são efeitos sazonais não meteorológicos, tais como feriados e férias escolares (Koetse e Rietveld, 2009). Pereira *et al.* (2014) mencionam ainda outros contextos como eventos especiais, cerimonias religiosas e obras. No geral, dentre outras informação de contexto, os incidentes, os eventos especiais, as obstruções na via e os dados meteorológicos são os que mais influenciam o tráfego, segundo Kwon *et al.* (2006). Eventos especiais planejados como shows, jogos de futebol e concertos são situações de contexto frequentemente estudadas. Kwoczek *et al.* (2014) descrevem uma ferramenta capaz de predizer a ocorrência de congestionamentos com base nesses eventos especiais e identificam segmentos rodoviários afetados por esses eventos (Kwoczek *et al.*, 2015).

Para relacionar informações de contexto às condições do tráfego são usados métodos estatísticos ou métodos de aprendizado de máquina, como classificação. Em ambos os casos, determinar se, por exemplo, a temperatura é um atributo importante dentre as informações meteorológicas é indispensável para obtenção de bons resultados. A seleção de atributos é um processo que determina a relação de importâncias entre valores de atributos e como uma informação (instância) foi classificada (Chandrashekar e Sahin, 2014).

Conhecer a importância dos atributos é fundamental para diversas tarefas, como por exemplo para a construção de árvores de decisão (AD). Uma AD lida com decisões, que, por sua vez, são baseadas nos atributos das instâncias do conjunto de dados em questão. Para ilustrar, usamos um exemplo simplificado onde uma pessoa está sendo classificada como “Em forma” ou “Fora de forma” com base em três diferentes atributos (idade, faz exercício pela manhã ou não, come muita pizza ou não). Há várias possibilidades de construir uma AD; porém tenta-se fazer isto

Tabela 1: Conjunto de dados de treino.

Idade (> 30)	**Como pizza?**	**Faz exercício?**	**Condição**
sim	não	sim	Fora de Forma
sim	sim	sim	Fora de Forma
não	não	não	Em Forma
não	sim	sim	Em Forma
sim	sim	não	Fora de Forma
não	não	sim	Em Forma
não	sim	não	Fora de Forma
sim	não	sim	Fora de Forma

de forma mais otimizada possível. Existem diversas métricas que permitem reconhecer quais atributos fornecem maior informação que outros, e para a construção de ADs é importante a escolha da uma métrica apropriada (isto pode depender da complexidade do conjunto de dados).

Com base em pesquisas sobre seleção de atributos (Guyon e Elisseeff, 2003), o ganho de informação (IG, *information gain*) foi selecionado como métrica para medir a importância de atributos. IG é uma métrica usada em métodos de seleção de atributos baseados em filtro, ou seja, métodos onde apenas são observadas as propriedades intrínsecas dos dados. Esta métrica é facilmente escalável para conjuntos de dados de dimensões muito altas, é computacionalmente simples e rápida, e é independente do modelo de classificação. Esta última característica é importante no escopo desta pesquisa, pois o principal interesse é na análise de atributos.

O IG detecta os atributos de maior importância através do cálculo da *entropia*, uma medida da pureza ou impureza de um determinado conjunto. A pergunta que ela responde é: quão diferentes/iguais os elementos são entre si? Quando definida sobre um conjunto de treinamento S, a entropia é calculada usando a Equação 1, onde $p(x)$ é a proporção de exemplos em relação a todo conjunto, e n é o número de rótulos disponíveis que estão sendo usados (no exemplo temos dois rótulos: "Em forma" ou "Fora de forma").

$$H(S) = \sum_{i=1}^{n} -p(x_i) \cdot \log_2 p(x_i) \tag{1}$$

Para ilustrar o cálculo da entropia, utilizamos o conjunto de dados apresentado na Tabela 1, que consiste de um conjunto de treinamento referente ao exemplo anteriormente mencionado, com 8 instâncias, tendo cada uma 3 atributos e um rótulo (última coluna). Na tabela temos 5 instâncias rotuladas como "Fora de forma" e 3 como "Em forma". A entropia é então $H(S) = -\frac{5}{8} \cdot \log_2(\frac{5}{8}) - \frac{3}{8} \cdot \log_2(\frac{3}{8}) = -(-0.531) - (-0.424) = 0.955$

O valor da entropia total do conjunto de dados S é usado para o cálculo do IG de cada atributo, e com isto, por exemplo, para selecionar o atributo mais indicado para ser o primeiro nó de uma AD. O IG de um atributo A é calculado usando a Equação 2, onde A representa o atributo em questão e $p(A_v)$ é a proporção da frequência de cada valor de A em relação ao conjunto todo.

$$IG(S, A) = H(S) - \sum_{v \in valores(A)} p(A_v) \cdot H(A_v) \tag{2}$$

Tabela 2: Tabela de Frequência para o atributo "idade".

Valores	Frequência relativa	Em forma	Fora de forma
sim	4	0	4
não	4	3	1

O IG nesse contexto é um índice que mede qual o melhor atributo para ramificar o primeiro nó da AD. Conforme visto, a entropia do conjunto todo vale 0.955. Com isso podemos calcular o IG do atributo "idade". A Tabela 2 mostra a frequência relativa de cada valor juntamente com a ocorrência de valores para cada um dos rótulos. Com isso podemos calcular o IG da primeira coluna e saber se ela é um bom atributo para ser a raiz da árvore. Então teremos: $IG(S, idade) = 0.955 - \frac{4}{8} \cdot H(sim) - \frac{4}{8} \cdot H(nao)$.

Calculando a entropia do conjunto para "idade" teremos: $IG(S, idade) = 0.955 - 0 - (0.406) = 0.549$. Realizamos o mesmo procedimento para os atributos "exercício pela manhã" e "come muita pizza": $IG(S, exercicio) = 0.215$ e $IG(S, comepizza) = 0.135$. Desta forma podemos concluir que o melhor atributo para ser a raiz da árvore seria o atributo "idade", porque esse atributo apresenta um maior valor de IG.

3. MATERIAIS E MÉTODOS

Nesta seção são descritos os conjuntos de dados e como se calcular o IG sobre eles.

3.1. Conjunto de Dados

Para relacionar contextos e condições do tráfego foram usados dois conjuntos de dados. Objetiva-se aqui verificar se e como contextos podem explicar (ou não) determinadas condições de tráfego.

O primeiro conjunto de dados é composto por todas as informações de contextos, ou seja, são informações, a priori, não relacionadas ao tráfego em si. Tais informações, em geral, possuem uma duração determinada e se referem a um local específico. As informações que compõem esse primeiro conjunto foram obtidas de: websites sobre condições meteorológicas; Facebook para extração dos eventos especiais e esportivos; Twitter para extração de informações de obras e manifestações e sites contendo calendários para extração de feriados e vésperas de feriados.

A Tabela 3 relaciona os tipos de contextos e seus atributos. O contexto relativo às condições meteorológicas (`https://openweathermap.org`), por exemplo, é composto por umidade, temperatura, velocidade do vento, pressão atmosférica, nível de precipitação e condição do tempo (nublado, etc.). Eventos especiais (`https://facebook.com/events`) são shows, encontros, concertos ou qualquer aglomerado de pessoas previamente agendado; também foram considerados eventos esportivos nos dois principais estádios de futebol: o Beira-Rio e a Arena do Grêmio. Manifestações (`https://twitter.com/TransitoPOARS`) são aglomerações de pessoas, onde normalmente há o bloqueio parcial ou completo da via onde se realiza. Contextos sobre obras (`https://twitter.com/TransitoPOARS`) são informações sobre a ocorrência de qualquer tipo de obra próxima a uma via ou na própria via. Foram considerados feriados (`http://calendario.com.br`) nacionais e estaduais no estado do Rio Grande do Sul, e suas vésperas. No total, são 7 tipos de informações de contexto e 16 atributos.

O segundo conjunto de dados é o que trata das condições do tráfego em diferentes dias, horários e locais na cidade de Porto Alegre. Essas informações são obtidos de uma conta oficial da

Figura 1: Tweet da EPTC - Porto Alegre: Aviso de congestionamento na Av. Assis Brasil.

Tabela 3: Atributos por contextos

Contexto	Atributos
(C_1) Condições Meteorológicas	(A_1) Umidade, (A_2) Percentual de nuvens, (A_3) Temperatura, (A_4) Velocidade do vento, (A_5) Pressão atmosférica, (A_6) Nível de precipitação e (A_7) Condição do tempo
(C_2) Eventos Especiais	(A_8) Shows, Encontros, Concertos e Festas
(C_3) Manifestações	(A_9) Greves, Protestos e Paradas
(C_4) Obras	(A_{10}) Construções, Reformas, Pavimentação e Recapeamento
(C_5) Feriados	(A_{11}) Feriados e (A_{12}) Vésperas de Feriados
(C_6) Data e Hora	(A_{13}) Dia da semana e (A_{14}) Hora do Dia
(C_7) Local	(A_{15}) Latitude e (A_{16}) Longitude

EPTC (Empresa Pública de Transporte e Circulação de Porto Alegre) no *Twitter* (`https://twitter.com/EPTC_POA`) e cada mensagem foi classificada manualmente considerando um ou mais rótulos descritos na primeira coluna da Tabela 4. Destacamos portanto que se trata de uma classificação multi-rótulo, onde cada *tweet* pode mencionar informações sobre mais de um rótulo. A escolha deste conjunto de rótulos foi baseada na proposta de Albuquerque *et al.* (2016), onde os autores classificam *tweets* relacionados ao tráfego. Porém, adaptamos o conjunto para as particularidades do domínio aqui tratado.

Tabela 4: Rótulos utilizados na classificação dos *tweets* da EPTC

Rótulos	Exemplos
(R_1) Incidentes	acidente, panes, atropelamentos
(R_2) Obstrução	elevação de ponte, bloqueios, desvios de tráfego
(R_3) Semáforo	problemas em controladores, intervenção humana
(R_4) Condições Meteorológicas	alagamentos
(R_5) Tráfego Livre	rotas livres, tráfego fluindo normalmente
(R_6) Tráfego Pesado	congestionamento, filas
(R_7) Outros	condições não classificadas

Os dois conjuntos de dados foram extraídos durante os meses de abril, maio e junho de 2018. Após a coleta e eventual classificação (rotulagem) manual (caso dos tweets da EPTC), todos os dados passam por várias etapas de processamento, o que inclui georreferenciamento, relacionamentos espaciais e a seleção de atributos em si. Este processo é mostrado na Figura 2.

Desta forma, após a extração dos dados, o passo seguinte é identificar a localização de cada

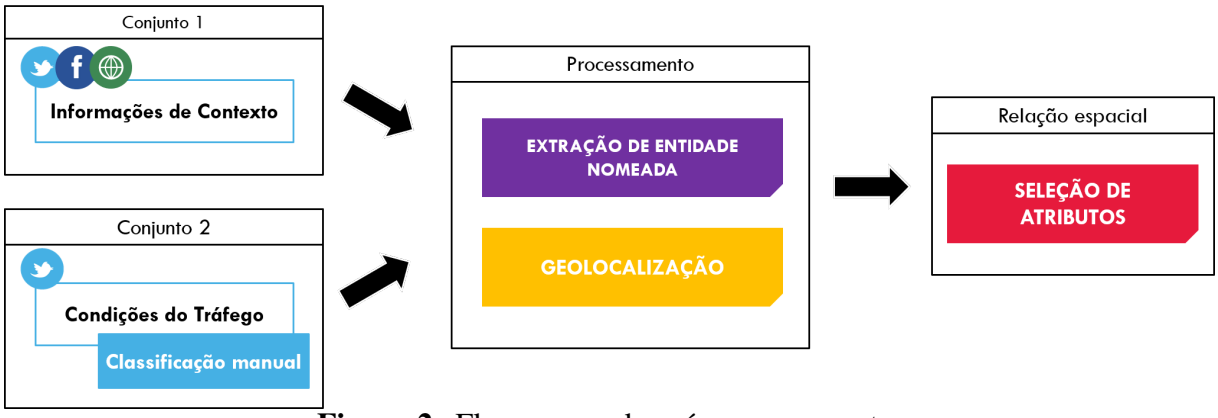

Figura 2: Fluxograma do pré processamento.

informação, tanto sobre contexto quanto sobre cada condição do tráfego. Esta tarefa não é trivial dado que envolve processamento de linguagem natural (mais especificamente, o que torna o processo ainda menos trivial, em língua portuguesa). Para tal, um modulo de extração de entidades nomeadas identifica nomes (de logradouros ou locais) que aparecem nos *tweets*. O modulo de extração de entidades nomeadas opera com a biblioteca Spacy (Honnibal e Montani, 2017), que implementa métodos estatísticos e de aprendizado de máquina para gerar um modelo capaz de localizar entidades nomeadas contidas em textos.

Após tal passo, pode-se proceder à etapa de georreferenciamento, que objetiva identificar a posição geográfica dos locais nomeados. Tal tarefa é igualmente não trivial pois raramente um *tweet* menciona o local exato relativo à informação. Em geral, um *tweet* refere-se a avenidas extensas. No caso da Figura 1, o *tweet* menciona duas delas, a Avenida Assis Brasil e a Avenida do Forte.

Na falta do georreferenciamento preciso, neste trabalho optou-se por discretizar a malha viária utilizando-se uma grade com um determinado tamanho de célula. A Figura 3 mostra tal esquema, onde as células marcadas (cor laranja) representam que uma informação de contexto ou condição do tráfego refere-se a um trecho da Avenida Assis Brasil e da Avenida do Forte. A Seção 4 mostra como o tamanho da célula afeta o desempenho da seleção de atributos e, consequentemente, da classificação.

Uma vez feita a etapa de georreferenciamento, a próxima etapa é relacionar as condição do tráfego e informações de contexto que ocorreram nas células afetadas. Por exemplo, caso ocorra um evento em uma célula, e nessa mesma célula seja reportado pela EPTC a condição de tráfego pesado, essas informações são relacionadas. Com isto, é possível investigar se a causa ou a justificativa da condição reportada é o contexto em questão.

3.2. Ganho de informação

Para cada conjunto de dados mediu-se a importância de cada atributo. Para isto foi utilizado o IG (descrito na Seção 2) como métrica para verificar o quão relevantes são os atributos para cada um dos rótulos listados na Tabela 4. A análise é feita para cada rótulo, ou seja, para cada rótulo, os valores de IG serão calculados para cada atributo. Dado que o IG representa o quanto um atributo ajuda a reduzir a incerteza no resultado da previsão, um valor próximo de 0 significa que o atributo possui pouca importância, enquanto que valores positivos quantificam o quanto

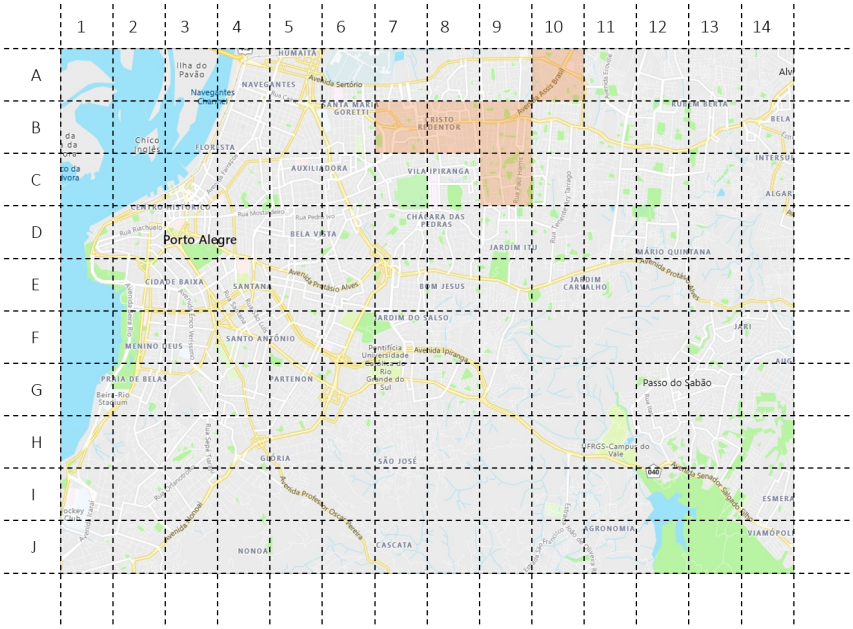

Figura 3: Mapa de Porto Alegre discretizado em grade.

um atributo consegue reduzir da incerteza do conjunto de dados.

4. EXPERIMENTOS E RESULTADOS

Como descrito na Subseção 3.1, os conjuntos de dados sobre informações de contextos (diversas fontes) e sobre condição do trânsito (*Twitter* da EPTC) precisam ser relacionados, e a relação é feito com base na associação da localização das informações com as células que constituem a discretização da malha viária. Uma vez que a associação é feita para todas as células afetadas, o tamanho da célula é uma variável importante neste processo. Foram utilizadas células de 50m x 50m, 200m x 200m e 500m x 500m, gerando três cenários distintos.

Para cada cenário, criamos um banco de dados. O tamanho destes bancos de dados (número de instâncias) variam em função da discretização do cenários. A primeira linha da Tabela 5 mostra o tamanhos de cada banco de dados. Por exemplo, a partir do *tweet* mostrado na Figura 1, vimos que, na Figura 3, cinco células são afetadas. Com isto, são gerados cinco instâncias que tratam dessa informação de contexto, uma para cada célula. Em cada uma destas instância, apenas o atributo referente à latitude e à longitude têm valores distintos. Este tipo de propagação espacial de uma informação vale para eventos especiais, manifestações e obras. Entretando, as informações sobre condições meteorológicas e feriados (e vésperas) são propagadas para todas as células, uma vez que elas valem para toda a malha. Notar ainda que há muito mais informações nas redes sociais em vésperas de feriados do que nos feriados em si.

A Tabela 5 mostra o número de instâncias geradas em cada cenário no seu respectivo banco de dados. A Tabela 6 descreve o percentual de rótulos por cenário.

Nas seções seguintes são apresentadas as conclusões das análises realizadas, tanto sobre a relevância dos atributos em relação a cada rótulo (dentre aqueles que aparecem na Tabela 4), quanto sobre o efeito da discretização da malha viária. No primeiro caso, a título de exemplo, é usado o cenário 2, onde a discretização envolve células de 200m x 200m.

Tabela 5: Número de instâncias por cenário

	Cenários		
	50 x 50	200 x 200	500 x 500
Total de instâncias de informação de contexto	425.237	40.254	13.067
Quantidade de instâncias que se referem a eventos especiais	743	211	219
Quantidade de instâncias que se referem a obras	621	179	131
Quantidade de instâncias que se referem a manifestações	6205	551	169
Quantidade de instâncias que se referem a feriados	471	110	61
Quantidade de instâncias que se referem a véspera de feriados	3.616	612	315

Tabela 6: Porcentagem de rótulos por cenário

	Cenário		
Rótulos	**50 x 50**	**200 x 200**	**500 x 500**
Incidentes	41,23%	35,4%	30,6%
Obstrução	11,6%	14,9%	18,3%
Semáforos	1,2%	2,5%	3%
Condições Meteorológicas	1,1%	1,4%	1,8%
Tráfego Livre	11,3%	12%	12,4%
Tráfego Pesado	79%	76,5%	73,5%
Outros	8,4%	8,1%	8,1%

Desta forma, a seguir, são apresentados os valores do IG relativo aos diversos rótulos (ver Tabela 4. As figuras 4 a 6 apresentam valores percentuais (destacando percentuais baixos que seriam pouco visíveis) para o caso de discretização intermediaria. Já as figuras 7 a 9 trazem os valores absolutos de IG para cada rótulo, em cada discretização.

4.1. Relevância dos atributos por rótulos

A Figura 4 mostra qual é a relevância percentual de cada um dos 16 atributos, para caso do rótulo "Tráfego Pesado". É possível observar que, para a condição de tráfego pesado, as condições meteorológicas (tons de azul), local (tons de verde) e horário (tons de amarelos) são bastante relevantes.

Em relação ao rótulo "Obstrução" (Figura 5), o panorama é similar, com leve aumento da importância dos atributos ligados às condições meteorológicas.

Já no caso do rótulo "Incidente" (Figura 6), as condições meteorológicas são ainda mais relevantes. Nota-se uma redução da relevância de atributos sobre manifestações, local e horário. Parece haver portanto uma baixa relação entre incidentes e estes atributos.

4.2. Relevância dos atributos em função dos cenários de discretização da malha viária

As Figuras 7, 8 e 9 descrevem a importância de cada atributo individualmente para cada rótulo nos cenários onde as células têm 50m x 50m, 200m x 200m e 500m x 500m respectivamente.

Individualmente, cada figura mostra a diferença de relevância entre os mesmo atributos e os diferentes rótulos. Na Figura 7 é possível observar, por exemplo, que local, dia da semana e horário não têm a mesma relevância absoluta quando se trata de determinar tráfego pesado e tráfego livre. No casos dos rótulos "Condições Meteorológica" e "Semáforos", todos os atributos possuem baixa relevância comparado aos demais rótulo. A condição de "Semáforos" trata,

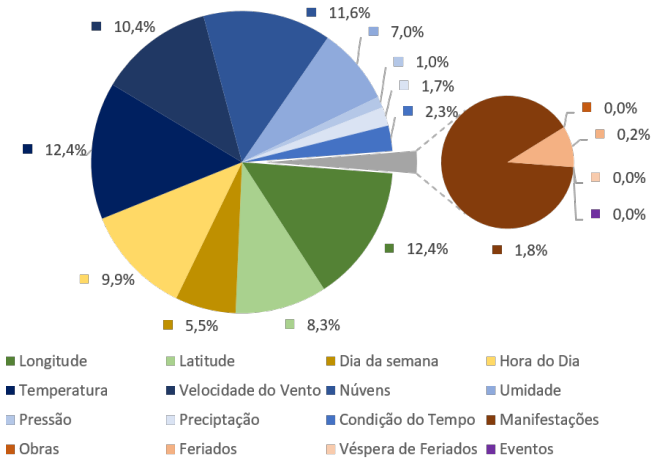

Figura 4: IG (percentual) dos atributos para o rótulo **Tráfego Pesado** (células de 200m x 200m)

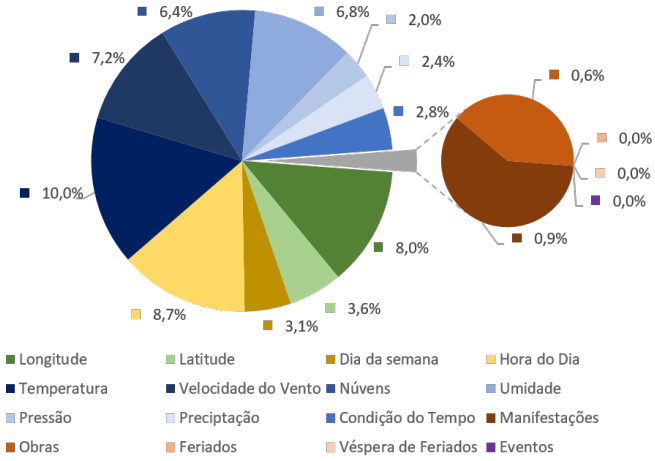

Figura 5: IG (percentual) dos atributos para o rótulo **Obstrução** (células de 200m x 200m)

principalmente, de falhas nos controladores semafóricos, ou seja, não é uma condição altamente dependente de informações de contextos externos. Além disto, como se nota na Tabela 6, há poucas instâncias que foram rotuladas para "Condições Meteorológica" e "Semáforos" pois apenas uma pequena parcela dos *tweets* menciona tais informações, o que tem efeito sobre a entropia. Isto vale para os três cenários.

No caso de "Condições Meteorológicas", este é o atributo com menor volume de informação nos três cenários. Houve por volta de quatro ocorrências de alagamento na cidade de Porto Alegre durante o período do estudo. Embora comparada aos outros rótulos, a importância dos atributos seja pequena, proporcionalmente é possível observar que há locais e horários mais propensos a alagamentos e outras condições e, evidentemente, há influência dos atributos meteorológicos sobre esse rótulo.

Ao comparar as três figuras (notar que os valores do IG no eixo x é diferente em cada uma delas), podemos concluir que o nível de discretização dos cenários pode alterar o valor da relevância dos atributos, mas não altera significativamente a ordem e porcentagem na relação entre atributos e rótulos.

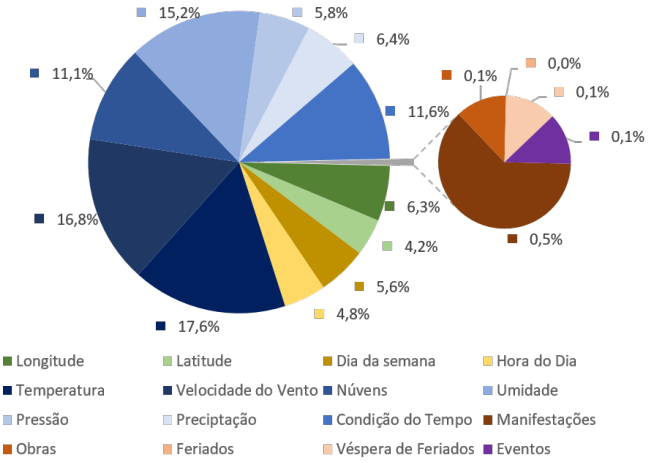

Figura 6: IG (percentual) dos atributos para o rótulo **Incidente** (células de 200m x 200m)

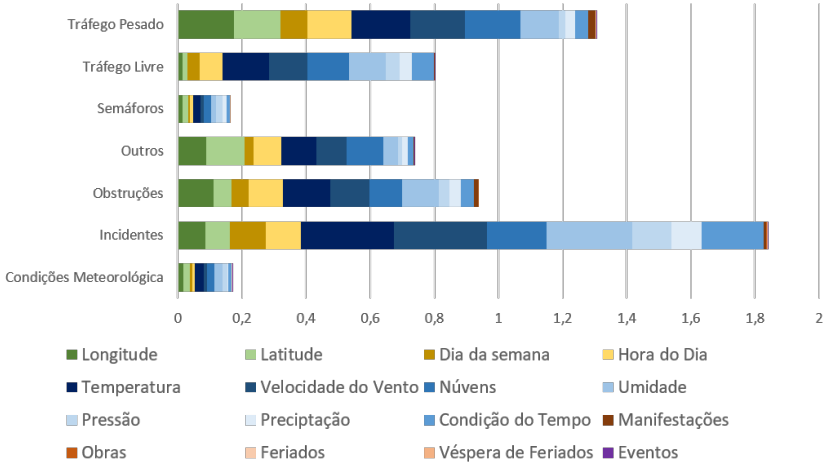

Figura 7: IG atributos x rótulos: Grade com células de 50m x 50m

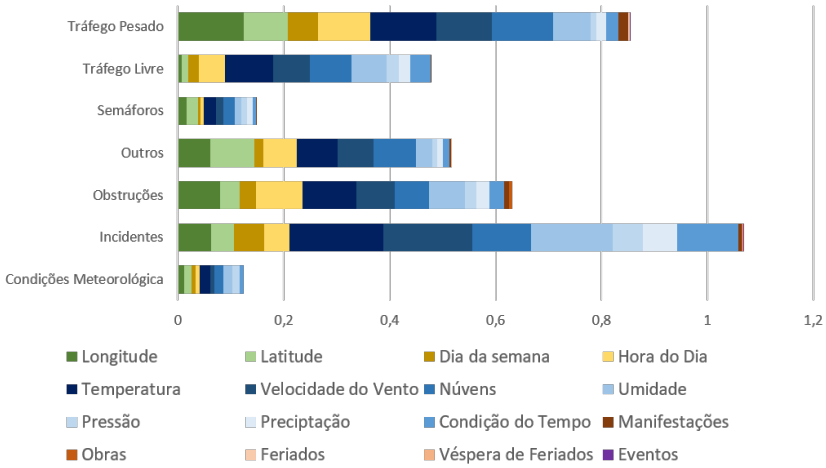

Figura 8: IG atributos x rótulos: Grade com células de 200m x 200m

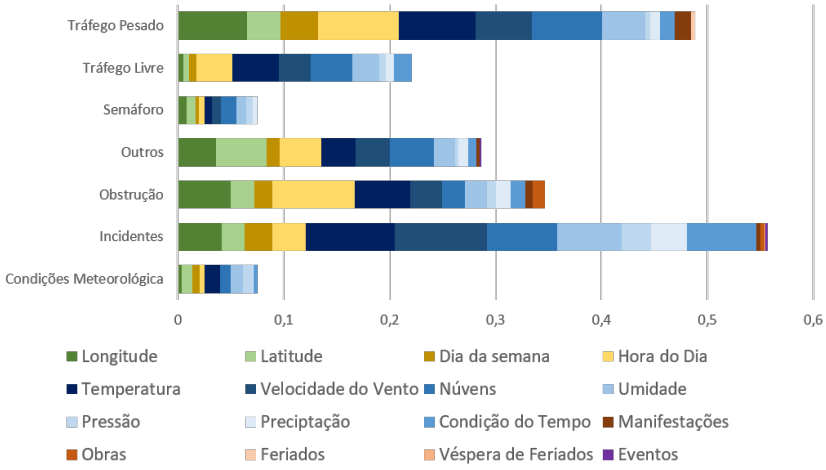

Figura 9: IG atributos x rótulos: Grade com células de 500m x 500m

5. CONCLUSÃO

Este trabalho objetiva relacionar informações de contexto com observações sobre tráfego, obtidas de fontes heterogêneas. Especificamente, foi feito um estudo da relevância de certos atributos para verificar a possibilidade de explicar determinadas observações.

Pode-se concluir, com os dados considerados aqui, que os atributos referentes às condições meteorológicas apresentam maior importância para a maioria dos rótulos, enquanto que localidade aparece em seguida. Os atributos "Dia da Semana" e "Hora do Dia" possuem importância significativa, principalmente para determinar condições de trânsito pesado e de obstrução. O atributo hora do dia é mais relevante que o dia da semana.

Trabalhos futuros incluem a análise da correlação entre os atributos. Por exemplo, no caso de condições meteorológicas, vários atributos são altamente correlacionados (como vento, pressão, temperatura, etc.). O método aqui empregado não levou em conta tal correlação. Além disto, as informações sobre condições meteorológicas foram propagadas igualmente para toda a malha viária. No caso de chuva, isto pode ser válido apenas para pequenos e médios municípios.

Agradecimentos

Este projeto é parcialmente financiado pelas agências MCTI/MC/CGI, Fundação de Amparo à Pesquisa do Estado de S. Paulo (FAPESP), sob número 2015/24423-3, e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior do Brasil (CAPES). Ana Bazzan agradece o apoio do CNPq (307215/2017-2).

REFERÊNCIAS

Albuquerque, F. C.; Casanova, M. A.; Lopes, H. *et al.* (2016) A Methodology for Traffic-related Twitter Messages Interpretation. *Computers in Industry*, vol. 78(C), 57–69.

Brodsky, H. e Hakkert, A. S. (1988) Risk of a road accident in rainy weather. *Accident Analysis & Prevention*, vol. 20(3), 161–176.

Chandrashekar, G. e Sahin, F. (2014) A survey on feature selection methods. *Computers & Electrical Engineering*, vol. 40(1), 16 – 28.

Edwards, J. B. (1996) Weather-related road accidents in England and Wales: a spatial analysis. *Journal of transport geography*, vol. 4(3), 201–212.

Eisenberg, D. (2004) The mixed effects of precipitation on traffic crashes. *Accident analysis & prevention*, vol. 36(4),

637–647.

Guyon, I. e Elisseeff, A. (2003) An introduction to variable and feature selection. *Journal of Machine Learning Research*, vol. 3, 1157–1182.

Honnibal, M. e Montani, I. (2017) spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. URL `https://github.com/explosion/spaCy`.

Ibrahim, A. T. e Hall, F. L. (1994) *Effect of adverse weather conditions on speed-flow-occupancy relationships*, p. 184–191. No. 1457 Em Transportation Research Record. Transportation Research Board. URL `https://trid.trb.org/view/425358`.

Koetse, M. J. e Rietveld, P. (2009) The impact of climate change and weather on transport: An overview of empirical findings. *Transportation Research Part D: Transport and Environment*, vol. 14(3), 205–221.

Kwoczek, S.; Di Martino, S. e Nejdl, W. (2014) Predicting and visualizing traffic congestion in the presence of planned special events. *Journal of Visual Languages & Computing*, vol. 25(6), 973–980.

Kwoczek, S.; Di Martino, S. e Nejdl, W. (2015) Stuck around the stadium? an approach to identify road segments affected by planned special events. Em: *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, p. 1255–1260. IEEE.

Kwon, J.; Mauch, M. e Varaiya, P. (2006) Components of congestion: Delay from incidents, special events, lane closures, weather, potential ramp metering gain, and excess demand. *Transportation Research Record*, vol. 1959(1), 84–91.

Maze, T. H.; Agarwal, M. e Burchett, G. (2006) Whether weather matters to traffic demand, traffic safety, and traffic operations and flow. *Transportation research record*, vol. 1948(1), 170–176.

Pereira, F. C.; Bazzan, A. L. C. e Ben-Akiva, M. (2014) The role of context in transport prediction. *IEEE Intelligent Systems Magazine*, vol. 29(1), 76–80. ITS Department.

Qiu, L. e Nixon, W. A. (2008) Effects of adverse weather on traffic crashes: systematic review and meta-analysis. *Transportation Research Record*, vol. 2055(1), 139–146.

Wang, S.; Djahel, S. e McManis, J. (2014) A Multi-Agent based vehicles re-routing system for unexpected traffic congestion avoidance. Em: *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, p. 2541–2548. IEEE.

Yazici, M. A.; Mudigonda, S. e Kamga, C. (2017) Incident Detection Through Twitter: Organization Versus Personal Accounts. *Transportation Research Record*, vol. 2643(1), 121–128.