PRIORIZAÇÃO DE VARIÁVEIS EXPLICATIVAS NA MODELAGEM DE ACIDENTES DE TRÂNSITO UTILIZANDO TÉCNICAS DE APRENDIZADO DE MÁQUINA

Philippe Barbosa Silva
Instituto Federal Goiano – Campus Rio Verde
Curso de Engenharia Civil
Universidade de Brasília
Programa de Pós-Graduação em Transportes
Michelle Andrade
Universidade de Brasília
Programa de Pós-Graduação em Transportes
Sara Ferreira
Universidade do Porto
Centro de Investigação do Território, Transportes e Ambiente

RESUMO
A priorização de variáveis no processo de modelagem de acidentes pode contribuir para otimização de recursos e indicação de quais dados são prioritários para coleta. Assim, este estudo objetivou investigar a influência da priorização de variáveis no ajuste de modelos de previsão de acidentes de resposta multivariada (número de acidentes sem vítimas, número de acidentes com vítimas e número de acidentes com mortes). Duas abordagens foram empregadas: técnicas de agrupamento de árvores de decisão (*Random Forest* e *Boosted Trees*) para a priorização inicial e posterior modelagem com uso de redes neurais artificiais (RNA); e, utilização direta de RNA para priorização e modelagem. Os resultados gerais, entretanto, indicaram piora no ajuste dos modelos quando da redução do número de variáveis explicativas. Apesar disso, acredita-se que a evolução de técnicas de aprendizado de máquina de dados que lidem melhor com resposta multivariada, conduzam à identificação adequada das variáveis mais importantes para modelagem.

ABSTRACT
The prioritization of variables in the process of accident modeling can contribute to the optimization of resource and indication of which data are priority to collect. This study aimed to investigate the influence of the prioritization of variables in the adjustment of accident predictive models of multivariate response (number of accidents without victims, number of accidents with victims and number of accidents with deaths). Two approaches were used: decision tree grouping techniques (Random Forest and Boosted Trees) for initial prioritization and later modeling using artificial neural networks (ANN); and, direct use of ANN for both, prioritization and modeling. Overall results, however, indicated a worse fitness of the models when the number of explanatory variables was reduced. Despite this, it is believed that the evolution of machine learning techniques that best deal with multivariate response, lead to the adequate identification of the most important variables for modeling.

1. INTRODUÇÃO

Para a abordagem de Sistemas Seguros, na medida em que os seres humanos cometem falhas, os projetistas de infraestrutura viária devem prover um sistema de transporte que minimize as consequências do erro humano. Para tanto, é essencial a investigação dos fatores contribuintes para a ocorrência dos acidentes de trânsito, onde a Modelagem da Segurança Viária (MSV) traz grande contribuição (Chang, 2005; Lord e Mannering, 2010; Cafiso *et al*., 2010; Costa *et al*., 2016).

Lord e Mannering (2010) reiteram que pesquisadores têm se dedicado à investigação dos fatores que afetam o número de acidentes que ocorre em algum espaço geográfico (normalmente uma interseção ou segmento de via) durante um período de tempo especificado, o que resulta em dados de frequência de acidentes e/ou severidade destes, objetivos dos modelos de previsão de acidentes (MPA).

Também tem ganhado força a abordagem de resposta multivariada para MPA. Tais modelos de previsão multivariada de frequência de acidentes consideram a interdependência no número de acidentes em diferentes níveis de severidade para um segmento rodoviário. El-Basyouny e Sayed (2009), Lee *et al*. (2015), Jonathan *et al*. (2016) e Ma *et al*. (2017), quando comparados os resultados, demonstraram a superioridade da abordagem de resposta multivariada frente aos modelos de uma variável-resposta.

A modelagem da segurança viária é tradicionalmente estatística, no entanto, mesmo diante dos avanços da modelagem estatística tradicional, são reconhecidas as limitações neste tipo de abordagem, uma vez que cada modelo estatístico tem pressupostos próprios e relação pré-definida entre variáveis dependentes e independentes (Zeng *et al*., 2016).

Mussone *et al*. (1999), Li *et al*. (2012) e Chang (2005) destacam que a modelagem estatística requer suposição sobre a distribuição dos dados e ainda, estabelece uma forma funcional entre variáveis dependentes e explicativas. Diversas vezes essas premissas podem não ser verdadeiras, e em sendo violadas, conduzem a estimativas equivocadas e produção incorreta de inferências. Os autores, reiterados por Abdelwahab e Abdel-Aty (2001), ainda evidenciam que o uso de redes neurais artificiais (RNA) não requer este tipo de relação pré-definida entre as variáveis. Nestes, em vez de elaborar uma forma funcional analítica, tarefa bastante complexa e laboriosa, é reconstruído um modelo, a partir do aprendizado realizado com os dados reais de acidentes, de onde obtém-se os pesos de cada variável do modelo. Hashemi *et al*. (1995) destacam que a RNA é capaz de identificar a relação entre os dados, enquanto a regressão requer conhecimento prévio da natureza do relacionamento subjacente.

A robustez e vantagem principal da utilização de RNA – ou outra técnica similar de Aprendizado de Máquina – está na sua capacidade de reconhecimento de padrões para detecção do relacionamento (linear ou não-linear) entre as variáveis. Tal aspecto que, conforme Abdelwahab e Abdel-Aty (2001), pode levar a uma maior compreensão da relação entre os fatores contribuintes e a ocorrência de acidentes. Além disso, Mannering e Bhat (2014) apontaram para a necessidade de exploração de modelos multivariados, mediante exploração de abordagens alternativas sugerindo, nomeadamente, a utilização de Aprendizado de Máquina (AM).

Ainda assim, uma limitação no processo de modelagem de acidentes é a obtenção de dados, uma vez que são necessários dados confiáveis para o ajuste dos modelos. Diversas vezes dados de variáveis não estão disponíveis nos bancos de dados, o que é justificado pelo custo de coleta e manutenção de dados detalhados ao longo de toda rede rodoviária, especialmente frente às limitações de recursos (financeiros, físicos e técnicos).

Conforme sugerido por Saha *et al*. (2015) e Saha *et al*. (2016), é relevante conduzir investigações sobre a possibilidade de simplificar os requisitos de dados a serem coletados, observando ainda a minimização dos impactos disso para a qualidade desejável dos modelos. Dessa forma, o objetivo deste trabalho é avaliar o impacto de priorização de variáveis no desempenho dos modelos ajustados. Tanto o processo de priorização quanto a modelagem foram procedidos com uso de técnicas de AM, empregando dados obtidos dos registros de acidentes entre os anos de 2011 e 2017 no trecho da BR-116 sob concessão da Nova Dutra.

2. MATERIAIS E MÉTODOS

2.1. Descrição dos dados

Os dados utilizados nesta investigação pertencem à pista norte da BR-116, que liga as cidades Rio de Janeiro e São Paulo. O trecho analisado é entre km 231,6 e km 0 no estado de São Paulo (SP) e km 333,5 e km 163 no estado do Rio de Janeiro (RJ), perfazendo um total aproximado de 402,1 km de extensão. O período analisado foi de 7 anos, de 2011 a 2017, mediante utilização do registro oficial de acidentes de trânsito fornecido pela Agência Nacional de Transportes Terrestres (ANTT). Na Tabela 1 estão apresentadas as estatísticas básicas da base de dados.

Tabela 1: Relação volume-velocidade medida no local

Ano	Número de acidentes sem vítimas (NASV)	Número de acidentes com vítimas (NACV)	Número de acidentes com mortes (NACM)	TOTAL
2011	3.452	1.273	55	4.780
2012	3.529	1.162	40	4.731
2013	3.359	1.149	47	4.555
2014	3.252	1.131	29	4.412
2015	2.764	1.041	23	3.828
2016	2.524	997	32	3.553
2017	2.491	952	37	3.480
TOTAL	21.371	7.705	263	29.339

2.2. Variáveis

O trecho analisado, após remoção de subtrechos com dados faltantes ou inconsistentes, teve a seguinte configuração: km 210 a km 0 (210 km) – trecho em São Paulo; km 333,5 a km 171 (162,5 km) – trecho no Rio de Janeiro; extensão total: 372,5 km. Tal trecho foi dividido em segmentos de extensão fixa de 500 m, totalizando 745 segmentos. E a partir dos dados disponíveis e dos achados provenientes da Revisão Sistemática de Literatura, foram criadas as seguintes variáveis explicativas para modelagem:

- Proporção do comprimento de reta no segmento (PCR) e Proporção do comprimento de curva no segmento (PCC) (Cafiso *et al*., 2010): adimensionais e descrevem a porcentagem de curva e reta no segmento, sendo sempre PCR + PCC = 1;
- Inverso do raio das curvas horizontais (IRH) (Costa *et al*., 2016): média dos inversos de raio das curvas horizontais presentes no segmento;
- Inclinação média das rampas ascendentes (IMA) e Inclinação média das rampas descendentes (IMD): descrevem o perfil vertical do segmento rodoviário, em referência à extensão de cada inclinação no segmento;
- Índice de oportunidade de acesso (IOA) e Índice de oportunidade de saída (IOS): descrevem a extensão disponível de faixa de mudança de velocidade em relação à velocidade regulamentada do trecho (Figura 1), calculados pelas Equações 1 e 2:

$$IOA = \sum_{i=1}^{n} \frac{l_{oa,\,i}}{V_{\max,\,i}} \tag{1}$$

$$IOS = \sum_{j=1}^{m} \frac{l_{os,\,j}}{V_{\max,\,j}} \tag{2}$$

em que $l_{oa,i}$ e $l_{os,j}$ são as extensões de faixa de mudança de velocidade para acesso e saída da rodovia nos sub-trechos i e j, respectivamente; $V_{max,i}$ e $V_{max,j}$ são as velocidades máximas regulamentadas para os sub-trechos i e j, respectivamente.

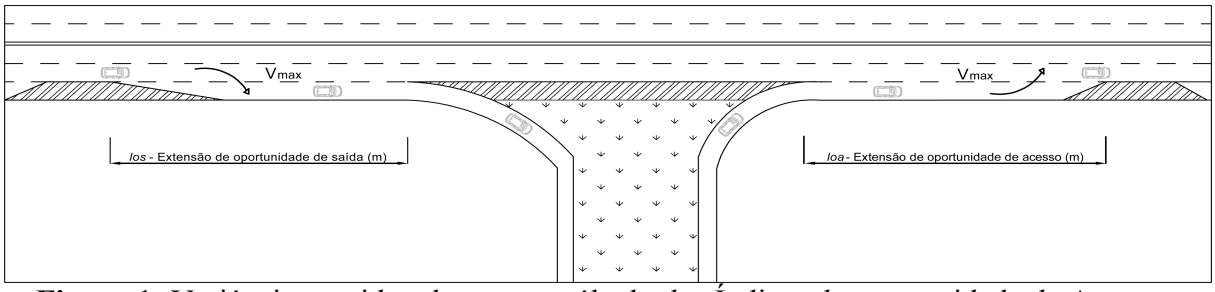

Figura 1: Variáveis consideradas para o cálculo dos Índices de oportunidade de Acesso e Saída

- Tipo de divisão de pista (Div_pista) divide-se em Div_pista_C para separação por canteiro central e Div_pista_B para divisão por meio de barreiras de concreto;
- L_cant: é definida como a largura do canteiro central;
- N_faixas: é definido como o número de faixas, por sentido, no segmento;
- L_acost: é definida como a largura do acostamento direito da pista;
- Volume de tráfego diário médio anual (VDMA): valor médio para o segmento;
- V_max: velocidade máxima regulamentada para o segmento;
- IS: índice de saturação do segmento, obtido pelo quociente do volume de tráfego pela capacidade do segmento;
- NS: nível de serviço (A, B, C, D ou E) do segmento;
- Cam_Pain: número de painéis fixos de mensagens variáveis e/ou câmeras no segmento;
- CEV: número de controladores eletrônicos de velocidade existente no segmento;
- N_Pass: número de travessias de pedestres existente no segmento;
- P_Ped: variável que indica a presença de praça de pedágio no segmento;
- Uso do solo (Uso_solo) divide-se em Uso_solo_R quando indica que a área envolvente do segmento é de uso rural ou Uso_solo_U quando a área envolvente do segmento é de uso urbano;
- P_ilum: proporção de iluminação artificial existente no segmento;
- QI: quociente médio de irregularidade do pavimento no segmento, medida normalizada no Brasil, similar ao IRI (International Roughness Index);
- IGG: índice de gravidade global do pavimento, no segmento considerado. Descreve o estado geral de um determinado trecho do pavimento;
- Def_max: valor médio da deflexão máxima do pavimento, no segmento.
- Para todos os segmentos foram utilizados os valores médios das variáveis (ou totais, quando o caso) à exceção dos Div_pista, N_faixas, V_max, Uso_Solo e NS. Para esses foi utilizada a característica predominante no segmento.

Na Tabela 2 estão apresentadas as estatísticas descritivas das variáveis para os segmentos, incluindo as três variáveis dependentes: Número de acidentes sem vítimas (NASV), Número de acidentes com vítimas (NACV) e Número de acidentes com mortes (NACM).

Tabela 2: Resumo das estatísticas descritivas das variáveis

Variável	Mín-Máx	Média	Desvio padrão
NASV	0-80	4,11	5,263
NACV	0-17	1,48	1,90
NACM	0-3	0,05	0,23
Variáveis explanatórias			
Numéricas		Categóricas	

Variável	Mín-Máx	Média	Desvio padrão		
VDMA [veic/dia]	11.052,85-79.550,16	22.633,15	10.956,18	Div_pista_B	3430 (65,80%)
IMA [%]	0,00-6,59	1,39	1,61	Div_pista_C	1785 (34,20%)
IMD [%]	0,00-38,00	1,71	2,33	NS_B	1622 (31,10%)
L_acost [m]	0,00-5,32	1,66	1,10	NS_C	3575 (68,56%)
L_cant [m]	0,00-2.000,00	47,49	298,41	NS_D	9 (0,17%)
N_faix	2-4	2,08	0,31	NS_E	9 (0,17%)
IOA [m/km/h]	0,00-8,97	0,61	1,55	P_Ped_Sim	77 (1,50%)
IOS [m/km/h]	0,00-12,44	0,56	1,17	P_Ped_Não	5138 (98,50%)
PCC	0,00-1,00	0,25	0,29	Uso_solo_Urb	3997 (76,60%)
PCR	0,00-1,00	0,75	0,29	Uso_solo_Rur	1218 (23,40%)
IRH [m^{-1}]	5,00E-05	1,22E-03	2,42E-03		
Variável	**Mín-Máx**	**Média**	**Desvio padrão**		
V_max [km/h]	40-110	99,25	14,28		
N_pass	0-1	0,07	0,26		
P_ilum	0,00-1,00	0,05	0,19		
CEV	0-1	0,04	0,19		
Cam_Pain	0-2	0,13	0,36		
IS	0,35-1,00	0,50	0,072		
IGG	1,02-13,22	5,24	2,11		
QI [cont/km]	23,23-33,08	27,91	2,60		
Def_max [0.01 mm]	18,91-42,66	31,43	5,04		

3. METODOLOGIA

Diversos estudos buscaram identificar e ranquear a influência das variáveis preditoras na previsão de acidentes, destacando-se o potencial de técnicas de árvores de decisão para tal finalidade. Na construção de AD é utilizado um conjunto de treinamento formado pelas entradas e saídas, estas últimas são as classes. A estrutura de uma AD contém um nó raiz (que inicia a árvore), nós de decisão (que dividem um determinado atributo e geram as ramificações) e folhas (que contém as informações de classificação). Cada nó indica o teste de um atributo, sendo a utilidade do atributo para a classificação utilizada como critério de ramificação. Dessa forma, o atributo escolhido, que será um nó da árvore, é aquele que gera maior ganho de informação (entropia), ou seja, melhor qualidade de classificação. Note-se que um percurso na árvore (da raiz à cada nó-folha) corresponde a uma regra de associação (Quinlan, 1986; Trabelsi *et al*., 2019). Os algoritmos de indução de árvores de decisão buscam, em meio a um conjunto de atributos, aqueles que separam da melhor forma os exemplos, gerando sub-árvores.

Nesta seção são apresentados os métodos *Random Forest* (RF) e *Boosted Trees* (BT), associação de árvores de decisão, empregados para a priorização das variáveis. Além disso, apresenta-se a descrição das RNA utilizadas para a modelagem.

3.1. *Random Forest*

Random Forest (RF), ou floresta aleatória, é uma técnica *ensemble* baseada em CART (Árvores de Classificação e Regressão). Conforme Dietterich (2000), *ensemble* é um conjunto de modelos que combinados produzem a predição de resposta para um novo caso. RF é um *ensemble* do tipo *bagging* (*bootstrap aggregating*), no qual várias predições de modelos independentes são agregadas.

Conforme proposto por Breiman (2001), as florestas aleatórias combinam árvores de classificação e/ou regressão, baseando-se em vetores de características (covariáveis), que são

geradas de maneira aleatória e independente a partir do conjunto original de dados.

Na Figura 2 está apresentado o esquema de funcionamento da RF. Formalmente é um classificador consistindo de uma coleção de árvores $\{h_k(x,\Theta_k), k=1,2,...,N\}$, em que Θ_k são vetores independentes e aleatoriamente distribuídos e cada árvore vota na classe mais popular para a entrada x (Breiman, 2001). As amostras são obtidas a partir do conjunto original de dados e, seguidamente, as árvores de classificação e/ou regressão são geradas com seleção aleatória das características (atributos) de cada amostra. Por fim, as árvores são combinadas para emitir a predição do conjunto. Quando classificador, a classe resultante é proveniente da maioria dos votos; nos regressores, o resultado é a média dos resultados de todas as árvores.

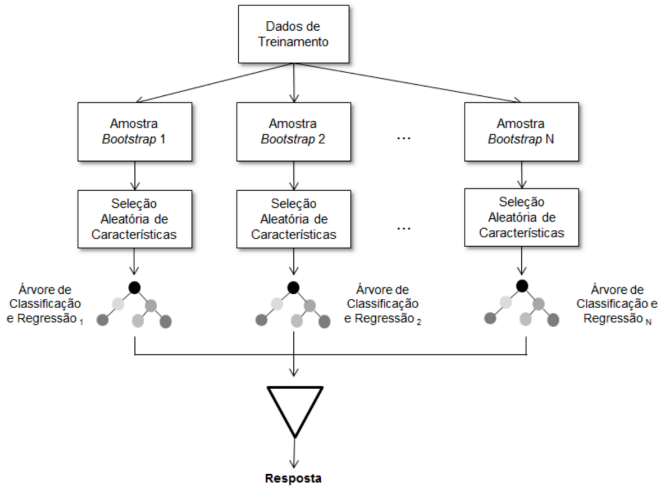

Figura 2: Esquema geral *Random Forest*
Fonte: Borges Junior (2016)

3.2. *Boosted Trees*

Boosting é outro método *ensemble*, reunindo predições de modelos independentes (classificadores ou regressores). No caso de *Boosted Trees* (BT), diferentes árvores simples são criadas reponderando sequencialmente os exemplos no conjunto de treinamento.

Conforme Wang *et al*. (2011), inicialmente, todos os exemplos são inicializados com pesos iguais. Cada exemplo classificado incorretamente pela árvore anterior recebe um peso maior na próxima iteração do treinamento, buscando classificá-lo corretamente. O erro é computado e os pesos são ajustados (reduzidos para exemplos classificados corretamente e aumentados para exemplos classificados erroneamente).

Desta forma, busca-se dar ênfase às observações mal ajustadas (observações que se desviam muito da média) com base nos resultados da árvore anterior. Previsões de muitos modelos fracos são combinadas, de modo a produzir uma forte previsão e melhorar a precisão do modelo (Bühlmann e Hothorn, 2007; De'ath, 2007).

O voto de cada classificador individual é ponderado proporcionalmente ao seu desempenho. Um dos algoritmos mais conhecidos é o *Gradient Boosting*, proposto por Friedman (2001). Ele se baseia na minimização de uma função de custo (perda) e utiliza o método de otimização de descida do gradiente.

3.3. Redes Neurais Artificiais

Uma RNA com múltiplas camadas (*multi layer perceptron* – MLP) é tipicamente composta por três tipos camadas: uma camada de entrada, uma camada de saída e uma ou mais camadas ocultas. A camada de entrada recebe os valores das variáveis explanatórias, ou seja, os dados de acidentes e características viário-ambientais. A camada oculta, composta por m neurônios, sumariza o peso dos valores de entrada das diferentes variáveis explanatórias e calcula os padrões de associação (Villiers e Barnard, 1993; Chang, 2005). Já a camada de saída, soma os valores dos diferentes neurônios ocultos e, na sequência, apresenta os valores de saída da rede, neste caso três variáveis de saída.

A arquitetura da rede utilizada é do tipo *feedforward*, caracterizada pela propagação dos sinais sempre das camadas anteriores para as posteriores. Em termos de treinamento, foi utilizado o algoritmo de retropropagação. Nele, conforme Haykin (2009), busca-se a minimização dos erros, a partir do ajuste dos pesos da rede em relação a um padrão de saída conhecido. Baseado no método de gradiente descendente, a função de custo está na direção e sentido em que a função tem taxa de variação mínima e garante que a rede caminhe na superfície em direção à maior redução do erro. Por fim, a função de ativação utilizada, relacionada à capacidade representativa das redes neurais e que introduz uma componente não linear, é do tipo tangente hiperbólica (*tanh* - $f(x)=(e^{x} - e^{-x})/(e^{x} + e^{-x})$), que se melhor ajustou frente a tradicional função sigmoide (*sigm* - $f(x)=1/(1+e^{-x})$), que é mais suscetível à saturação. Foi adotado o particionamento aleatório de 70% dos dados para treinamento e 30% para teste.

3.4. Ajustes para aplicação das técnicas

Uma limitação das técnicas RF e BT é o fato de lidarem apenas com um output. Tal fato implicou na decomposição do problema nas três respostas consideradas: número de acidentes sem vítimas (NASV), número de acidentes com vítimas (NACV) e número de acidentes com vítimas fatais (NACM). Além disso, buscando simular o efeito conjunto das três saídas, minimizar o problema de desbalanceamento do NACV e NACM - especialmente este último – e melhorar o desempenho do modelo, estabeleceu-se também como variável resposta a UPS (Unidade Padrão de Severidade) do segmento.

A UPS, instituída por DENATRAN (1987), foi escolhida como medida do grau de severidade do segmento por expressar o número de ocorrência por severidade, pela atribuição de peso a cada severidade, conforme Equação 3.

$$UPS = 1.NASV + 5.NACV + 13.NACM \quad (3)$$

Se refere que as variáveis categóricas passaram por uma transformação e foram convertidas em variáveis binárias. Dessa forma, a variável Divisão de pista (Div_pista) foi transformada em Div_pista_B e Div_pista_C; Nível de Serviço (NS) em NS_B, NS_C, NS_D e NS_E; a variável e Uso do solo (Uso_solo) em Uso_solo_U e Uso_solo_R.

A utilização do módulo AutoModel do *software* RapidMiner permitiu o uso de RF e BT com otimização, via algoritmos genéticos com 11 replicações, dos parâmetros para ambas as técnicas utilizadas. Finalizada a otimização dos parâmetros e obtenção dos melhores modelos, são apresentadas as medidas de desempenho dos modelos. Como medida de desempenho e de comparação entre modelos utilizou-se o erro relativo (ER), conforme Equação 4.

$$ER = \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (4)$$

em que y_i é o valor observado, $\hat{y_i}$ é o valor previsto, $\bar{y}_i$ é a média dos valores observados e n é o conjunto de casos.

Além disso, já com os modelos resultantes, o AutoModel permite identificar o fator de importância de cada variável de entrada para a saída do modelo. Essa importância local de um atributo é baseada em sua correlação com as predições na vizinhança do *input* selecionado. Assim, quanto mais o valor se aproxima de 1, mais forte é a associação entre o atributo e a variável resposta, ao passo que, valores negativos indicam relação do input com a saída, mas contradizendo o valor da predição (RapidMiner, 2018).

Em termos da série histórica dos dados, considerando que houve uma mudança no procedimento de registro de acidentes a partir de 2015, mediante operacionalização do sistema DAT (Declaração de acidente de trânsito) da Polícia Rodoviária Federal (PRF), preferiu-se investigar, além do período completo, os modelos ajustados para 2011 a 2014 e 2015 a 2017, separadamente.

4. RESULTADOS E DISCUSSÃO

Para cada agregação temporal (2011-2014, 2015-2017 e 2011-2017) e por cada técnica (*Random Forest* e *Boosted Trees*), foram desenvolvidos quatro modelos (NASV, NACV, NACM e UPS), resultando em 48 modelos. Na Tabela 3 estão apresentados os resultados dos fatores importantes dos modelos do período 2011-2017 para ambas as técnicas.

Tabela 3: Fatores importantes para predição – 2011-2017

Acidente sem vítima		**Acidente com vítima**		**Acidente com morte**		**UPS**	
Random Forest		*Random Forest*		*Random Forest*		*Random Forest*	
Fator	Import.	Fator	Import.	Fator	Import.	Fator	Import.
Inv_raio	0,36	IOA	0,38	Inv_raio	0,35	VDMA	0,35
VDMA	0,32	VDMA	0,32	IOA	0,28	Inv_raio	0,24
IOA	0,19	Inv_raio	0,26	Ind_sat	-0,27	QI	-0,23
Ind_sat	-0,17	Ind_sat	-0,24	NS_D	-0,19	IOA	0,21
Def_max	-0,12	QI	-0,23	IGG	0,13	Def_max	-0,18
QI	-0,12	Def_max	-0,11	IOS	0,11	Ind_sat	-0,15
IOS	0,07	Div_pista_B	-0,08	Incl_neg	0,11	IOS	0,11
RMSE	2,85	RMSE	1,412	RMSE	0,211	RMSE	9,044
R^2	0,521	R^2	0,402	R^2	0,024	R^2	0,461
Boosted Trees		*Boosted Trees*		*Boosted Trees*		*Boosted Trees*	
Fator	Import.	Fator	Import.	Fator	Import.	Fator	Import.
Inv_raio	0,43	Inv_raio	0,41	IOA	0,51	Inv_raio	0,5
VDMA	0,24	VDMA	0,41	Ind_sat	-0,32	VDMA	0,37
Ind_sat	0,11	IOA	0,22	Inv_raio	0,15	IOA	0,33
IOA	0,09	Div_pista_B	-0,08	Def_max	-0,05	IOS	0,11
IOS	0,09	Vel_max	0,07	Incl_pos	-0,03	Incl_pos	0,08
NS_D	-0,08	IOS	0,07	L_acost	0,03	P_ilum	0,07
Vel_max	0,07	Ind_sat	-0,06	IOS	0,03	Div_pista_B	-0,07
RMSE	0,481	RMSE	1,279	RMSE	0,213	RMSE	7,497
R^2	0,635	R^2	0,449	R^2	0,031	R^2	0,539

*RMSE: Raiz do erro quadrático médio; R^2: Coeficiente de determinação

Nota-se, como é suposto, que a depender da técnica e da variável de saída, os fatores mais importantes são diferentes. Tendo isso em conta e visando tirar o máximo proveito dos resultados, decidiu-se por uma análise tipo *ensemble*, utilizando todas as variáveis que figurassem pelo menos uma vez no conjunto de resultados. Adicionalmente, foi estabelecido como critério para o ranqueamento das variáveis o somatório do produto entre a importância do fator (em módulo) e o coeficiente de determinação do modelo.

Na Tabela 4 estão apresentadas as listas de variáveis selecionadas (na ordem decrescente de importância) para comporem a modelagem com RNA. Notou-se, inclusive, que apesar das diferentes técnicas e agregações temporais, VDMA, IRH e IOA sempre figuraram entre as variáveis mais importantes.

Tabela 4: Variáveis selecionadas para modelagem

2011-2014	IRH / VDMA / IOA / Def_max / IOS / IGG / IMD / P_ilum / QI / IS / L_cant / L_acost / Div_pista_C / Div_pista_B / NS_B / PCC / IMA
2015-2017	IRH / VMDA / IOA / QI / IOS / IS / P_ilum / V_max / IMD / L_cant / Uso_solo_U / IGG / IMA / Div_pista_B / Def_max / L_acost
2011-2017	IRH / VDMA / IOA / IS / QI / IOS / Def_max / Div_pista_B / V_max / NS_D / IMA / P_ilum / IGG / IMD / L_acost

De posse dos resultados da priorização de variáveis obtidos por meio da associação entre *Random Forest* e *Boosted Trees*, iniciou-se o processo de modelagem com redes neurais. Além disso, o fator de importância de cada *input* foi considerado no processo sucessivo de exclusão de variáveis, excluindo a menos importante a cada nova rodada até o mínimo de duas variáveis. Os resultados estão apresentados na Tabela 5. Os valores negritados sinalizam o menor valor (entre todos os modelos baseados nas variáveis priorizadas) de determinado tipo de erro.

Tabela 5: Resultados modelos resultantes do uso de variáveis priorizadas

	Número de variáveis						
	Erros: Treinamento/Teste						
2011-2014	**28**	**19**	**18**	**17**	**16**	**15**	**14**
ER médio	0,713/0,767	0,756/0,79	0,794/0,833	0,767/0,803	0,769/0,798	**0,751/0,785**	0,757/0,794
ER NASV	0,577/0,627	0,60/0,691	0,644/0,752	0,614/0,705	0,621/0,705	**0,585/0,675**	0,601/0,693
ER NACV	0,637/0,701	0,707/0,699	0,754/0,755	0,717/0,712	0,713/0,709	**0,689/0,688**	0,693/0,698
ER NACM	0,984/1,004	0,962/0,987	0,984/0,983	0,97/0,981	0,972/0,97	**0,978/0,977**	0,977/0,978
2015-2017	**28**	**18**	**17**	**16**	**15**	**14**	**13**
ER médio	0,742/0,773	0,824/0,844	0,825/0,845	0,821/0,834	0,816/0,83	0,802/0,829	**0,797/0,829**
ER NASV	0,461/0,655	0,744/0,699	0,741/0,711	0,745/0,684	0,746/0,69	0,714/0,659	**0,697/0,647**
ER NACV	0,649/0,685	0,74/0,713	0,753/0,72	0,737/0,71	0,728/0,714	0,719/0,709	**0,722/0,715**
ER NACM	0,986/0,984	0,988/1,02	0,979/1,01	0,981/1,0	0,974/0,99	0,974/1,01	**0,972/1,0**
2011-2017	**29**	**18**	**17**	**16**	**15**	**14**	**13**
ER médio	0,741/0,777	**0,768/0,816**	0,787/0,828	0,824/0,852	0,766/0,822	0,823/0,843	0,796/0,825
ER NASV	0,612/0,668	**0,618/0,712**	0,653/0,722	0,736/0,762	0,632/0,719	0,72/0,737	0,67/0,713
ER NACV	0,673/0,715	**0,707/0,72**	0,726/0,741	0,753/0,773	0,718/0,728	0,756/0,762	0,732/0,738
ER NACM	0,987/0,995	**0,978/0,981**	0,982/0,986	0,983/0,99	0,976/0,984	0,993/0,995	0,985/0,988

O primeiro modelo apresentado é o geral contendo todas as variáveis. Esse serviu de parâmetro para avaliar os modelos baseados nas variáveis priorizadas. Os resultados não foram animadores, uma vez que o desempenho dos melhores modelos se aproximou bastante

do modelo geral (sem redução inicial de variáveis). E mais, em todos os períodos (2011-2014, 2015-2017 e 2011-2017), os melhores modelos (destacados em negrito) tiveram desempenho inferior ao modelo geral com todas as variáveis de cada período.

Diante de tais resultados e, valendo-se da afirmação de Nisbet *et al.* (2018), os quais dizem ser interessante utilizar os fatores de importância das variáveis de entrada (provenientes da análise de sensibilidade) como estratégia para determinar o melhor conjunto de variáveis a serem incluídas em um modelo, decidiu-se por proceder a modelagem com RNA a partir de todas as variáveis, executando-se sucessivas exclusões de variáveis. Dessa forma, os inputs com menores fatores de importância, de dois a dois (até 5% de importância) ou um a um (importância superior a 5%), foram excluídos até o último modelo com apenas duas variáveis independentes. Assim, mediante análise das medidas de desempenhos dos modelos foram obtidas as melhores configurações.

Esse procedimento, além de determinar o melhor conjunto de variáveis, também já consiste na modelagem inicial do problema, possibilitando análises e discussões acerca dos resultados. Na Tabela 6 estão apresentados os melhores modelos para cada agregação temporal, a partir da comparação dos erros.

Tabela 6: Resultados modelos resultantes do uso de variáveis priorizadas

	2011-2017 (25 variáveis)	**2011-2014 (24 variáveis)**	**2015-2017 (14 variáveis)**
	Erros: Treinamento/Teste		
ER NASV	0,555/0,591	0,527/0,600	0,423/0,643
ER NACV	0,646/0,697	0,601/0,704	0,641/0,678
ER NACM	0,986/0,991	0,970/1,010	0,983/0,967
ER médio	0,729/0,76	0,70/0,771	0,682/0,763
RMSE NASV	3,902/4,086	4,099/4,632	2,986/3,284
RMSE NACV	1,526/1,587	1,555/1,720	1,393/1,367
RMSE NACM	0,228/0,229	0,239/0,261	0,210/0,176
RMSE médio	1,886/1,967	1,964/2,205	1,530/1,609
MAD NASV	2,313/2,563	2,433/2,849	1,960/2,094
MAD NACV	1,079/1,137	1,092/1,203	1,027/1,014
MAD NACM	0,096/0,097	0,101/0,111	0,083/0,073
MAD médio	1,163/1,266	1,209/1,388	1,023/1,061

*ER: Erro relativo; RMSE: Raiz do erro quadrático médio; MAD: Desvio médio absoluto

Os erros e a análise de resíduos permitiram verificar que os modelos para os períodos de 2011 a 2014 e 2011 a 2017 reforçaram o constado anteriormente, apontando para a melhoria de ajuste do modelo quando se utiliza um número maior de variáveis (24 ou 25 variáveis), próximo ao número total de variáveis disponíveis inicialmente. Para o período de 2015 a 2017, no entanto, verificou-se que o modelo com 14 variáveis teve melhor ajuste, embora o modelo com segundo melhor desempenho para aquele período tenha sido obtido com o uso de 26 variáveis. Não se tem, portanto, consistência para indicação de coleta de dados simplificada que assegure ajuste dos modelos próximo ao melhor desempenho possível.

4. CONSIDERAÇÕES FINAIS

O objetivo deste trabalho foi investigar a influência da redução do número de variáveis explicativas no processo de modelagem de acidentes em uma rodovia brasileira. A rodovia foi

segmentada em trechos de 500 m, sendo cada segmento caracterizado pelos valores médios ou predominantes das variáveis explanatórias selecionadas, que contemplaram características geométricas, operacionais e do pavimento da via. Os modelos foram estimados para uma amostra de trechos de rodovia federal brasileira, utilizando dados dos anos de 2011 a 2017.

A redução de dimensionalidade do problema com uso de técnicas de agrupamento de árvores de decisão (*Random Forest* e *Boosted Trees*) foi conduzida. Os resultados, no entanto, não foram satisfatórios, uma vez que houve decréscimo do desempenho do modelo quando utilizadas apenas as variáveis priorizadas. O fato de ter sido necessário decompor o problema multivariado em quatro abordagens univariadas e, os resultados destas análises terem sido conjuntamente compilados e utilizados como entrada numa abordagem multivariada é um possível motivador do desempenho insatisfatório. Ademais, a distinção do funcionamento das árvores de decisão e das redes neurais, por si só, pode ter influenciado nos resultados.

Para RNA, os resultados também sugerem que os melhores modelos são obtidos com uso de quase totalidade das variáveis. Como os conjuntos de variáveis resultantes da priorização continham entre 18 e 19 variáveis, comparativamente, pode-se dizer que essa configuração já estaria a jusante da configuração "ótima", confirmando os resultados inferiores.

Tanto na abordagem com RF e BT quanto na utilização direta de RNA para priorização de variáveis, não fica evidente qual a redução de variáveis poderia ser realizada preservando-se ao mesmo tempo o desempenho próximo ao melhor possível. Diferente de modelos estatísticos, as técnicas de Mineração de Dados identificam padrões não explícitos, o que pode sugerir que quanto mais dados (e informações) são utilizadas, melhor é a descoberta de conhecimento do problema (ajuste do modelo).

Apesar dos resultados, acredita-se que a evolução computacional de métodos de agrupamento de árvores de decisão para respostas multivariadas e aprimoramentos na determinação de fatores de importância nas RNA podem conduzir à melhoria dos resultados e permitir a priorização de variáveis, sem perda de desempenho. Além disso, acredita-se que, em uma abordagem de análise de acidentes por severidade, e não frequência de acidentes como nesta pesquisa, em que existem dados desagregados e em quantidade muito superior, tais técnicas podem ser mais eficazes.

Agradecimentos
O primeiro autor agradece ao apoio financeiro do Instituto Federal Goiano (IF Goiano). Refere-se ainda que o presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001 e da Fundação para a Ciência e Tecnologia- Portugal - (FCT) por meio do projeto "Mobilidade Urbana Sustentável e Segura", no qual este trabalho está inserido. Os autores também agradecem à ANTT (Agência Nacional de Transportes Terrestres) pela disponibilização dos dados.

REFERÊNCIAS BIBLIOGRÁFICAS

Abdelwahab, H.; Abdel-Aty, M. (2001) Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections. *Transportation Research Record* 1746, 6-13.

Borges Junior, S. R. (2016) *SEnsembles – uma abordagem para melhorar a qualidade das correspondências de instâncias disjuntas em estudos observacionais explorando características idênticas e ensembles de regressores*. Tese de Doutorado em Ciência da Computação, Universidade Federal de São Carlos, São Carlos.

Bühlmann, P.,; Hothorn, T. (2007) Boosting algorithms: regularization, prediction, and model fitting. *Statistical Science* 22(4), 477–505.

Cafiso, S.; Di Graziano, A.; Di Silvestro, G.; La Cavaa, G.; Persaud, B. (2010) Development of comprehensive accident models for two-lane rural highways using exposure, geometry, consistency and context variables. *Accident Analysis and Prevention* 42, 1072-1079.

Chang, L. (2005) Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. *Safety Science* 43, 541-557.

Costa, J. O.; Jacques, M. A. P.; Soares, F. E. C., Freitas, E. F. (2016) Integration of geometric consistency contributory factors in three-leg junctions collision prediction models of Portuguese two-lane national highways. *Accident Analysis and Prevention* 86, 59-67.

DENATRAN (1987) *Manual de identificação, análise e tratamento de pontos negros*, 2ª edição. Departamento Nacional de Trânsito, Brasília.

De'ath, G. (2007) Boosted trees for ecological modeling and prediction. *Ecology* 88(1), 243–251.

Dietterich, T. G. (2000) Ensemble methods in machine learning. *Proceedings of International Workshop on Multiple Classifier Systems*, Cagliari, Itália, p. 1–15.

El-Basyouny, K.; Sayed, T. (2009) Collision prediction models using multivariate Poisson-lognormal regression. *Accident Analysis and Prevention* 41(4), 820–828.

Friedman, J. H. (2001) Greedy function approximation: a gradient boosting machine. *The Annal of Statistcs* 29(5), 1189-1232.

Hashemi, R. L.; Blanc C. R.; Shearry, A. (1995) Neural Network for Transportation Safety Modeling. *Expert Systems with Applications* 9, 247–256.

Jonathan, A.; Wu, V.; Donnell, K.F.K.; Donnell, E.T. (2016) A multivariate spatialcrash frequency model for identifying sites with promise based on crash types. *Accident Analysis and Prevention* 87, 8–16.

Lee, J., Abdel-Aty, M., Jiang, X., (2015) Multivariate crash modeling for motor vehicleand non-motorized modes at the macroscopic level. Accid. *Accident Analysis and Prevention* 78,146–154.

Li, H.; Graham, D. J.; Majumdar, A. (2012) The effects of congestion charging on road traffic casualties: A causal analysis using difference-in-difference estimation. *Accident Analysis and Prevention* 49, 366-377.

Lord, D.; Mannering, F. (2010) The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A* 44, 291–305.

Ma, X.; Chen, S.; Chen, F. (2017) Multivariate space-time modeling of crash frequencies by injury severity levels. *Analytic Methods in Accident Research* 15, 29-40.

Mannering, F. L.; Bhat, C. R. (2014) Analytic Methods in Accident Research: Methodological Frontier and Future Directions. *Analytic Methods in Accident Research* 1, 1-22.

Mussone, L.; Ferrari, A.; Oneta, M. (1999) An analysis of urban collisions using an artificial intelligence model. *Accident Analysis and Prevention* 31(6), 705–718.

Quinlan, J. R. (1986) Induction of Decision Trees. *Machine Learning* 1, 81-106.

RapidMiner (2018) Explain Predictions (Model Simulator). Disponível em: https://docs.rapidminer.com/8.2/studio/operators/scoring/explain_predictions.html. Acesso em 20 out. 2018.

Saha, D.; Alluri, P.; Gan, A. (2015) Prioritizing Highway Safety Manual's crash prediction variables using boosted regression trees. *Accident Analysis and Prevention* 79, 133-144.

Saha, D.; Alluri, P.; Gan, A. (2016) A random forests approach to prioritize Highway Safety Manual (HSM) variables for data collection. *Journal of Advanced Transportation* 50, 522-540.

Trabelsi, A.; Elouedi, Z.; Lefevre, E. (2019) Decision tree classifiers for evidential atribute values nad class labels. *Fuzzy Sets and Systems* 366, 46-62.

Villiers, J.; Barnard, E. (1993) Back Backpropagation neural nets with one and two hidden layers. *IEEE Trans. Neural Netw.* 4(1), 136–141.

Wang, C.; Quddus, M. A.; Ison, S. G. (2011) Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. *Accident Analysis and Prevention* 43(6), 1979-1990.

Zeng, Q.; Huang, H.; Pei, X.; Wong, S. C. (2016) Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks. *Analytic Methods in Accident Research* 10, 12–25.

Philippe Barbosa Silva (philippe.silva@ifgoiano.edu.br)
Michelle Andrade (michelleandrade@unb.br)
Programa de Pós-Graduação em Transportes, Universidade de Brasília
Campus Darcy Ribeiro – Brasília, DF, Brasil
Sara Ferreira (sara@fe.up.pt)
Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias, s/ n – Porto, Portugal