DETERMINANTES E PREDIÇÃO DE CRIMES DE HOMICÍDIOS NO BRASIL: UMA ABORDAGEM DE APRENDIZADO DE MÁQUINA

Lucas Pereira LOPES[1]
Sabrina Vieira FELIX[2]

- RESUMO: Ao longo da história, as sociedades organizadas tentaram prevenir o crime seguindo várias abordagens, sendo que a justificativa social em entender quais características associadas à criminalidade é o seu uso para alcançar políticas públicas eficazes contra essas atividades ilegais. Nesse contexto, o objetivo deste trabalho é identificar os determinantes econômicos, sociais e demográficos dos crimes de homicídios no Brasil. E como objetivo secundário, realizar a predição do nível de crimes em território nacional. Como metodologia, trata-se de um estudo quantitativo onde se utilizou métodos de Árvore de Regressão, Florestas Aleatórias, *Boosting* e *K-Nearest Neighbors* (K-NN), como ferramentas alternativas aos modelos tradicionais lineares, como a regressão via mínimos quadrados ordinários. Os dados analisados indicaram que entre as 31 covariáveis utilizadas, 9 apresentaram os maiores impactos na violência em nível nacional, sendo, em ordem decrescente: o tamanho da população jovem, saneamento básico, tamanho da população total, população economicamente ativa, população urbana, PIB, mulheres chefes na família, pessoas pobres entre 0 e 14 anos e proporção de pessoas que ganham até meio salário mínimo, onde cada fator foi discutido de acordo com a literatura econômica do crime. Além disso, o modelo de Florestas Aleatórias explicou, em média, 82% da variabilidade dos crimes de homicídios em nível nacional. Acreditamos que esta abordagem ajuda a produzir respostas mais robustas sobre os efeitos dos fatores sociais, econômicos e demográficos sobre o crime, sendo, portanto, uma nova ferramenta para orientar os formuladores de políticas públicas.

- PALAVRAS-CHAVE: Florestas aleatórias; políticas públicas; violência; fatores socioeconômicos.

1 Introdução

A busca por determinantes de crimes de homicídios tem se tornado um dos grandes desafios na ciência do século XXI. Devido a quantidade de dados disponíveis (IPEA, 2018; IBGE, 2018; UCR, 2018; EUROSTAT, 2018; DATAGOVIN, 2018) e a facilidade aos seus acessos, pesquisadores sociais buscam entender quais variáveis impactam, e em qual magnitude, os níveis de violências nos grandes centros urbanos. Ao longo da história, as sociedades organizadas tentaram prevenir o crime seguindo várias abordagens (GORDON *et al*., 2009), sendo que a justificativa social em entender quais características

[1] Universidade de São Paulo - USP e Universidade Federal de São Carlos - UFSCar, CEP: 13566-590, São Carlos, SP, Brasil. E-mail: *luccas.lopes@usp.br*
[2] Universidade Federal de São Carlos - UFSCar, CEP: 13566-590, São Carlos, SP, Brasil. E-mail: *sabrinafelix@outlook.com*

associadas à criminalidade é o seu uso para alcançar políticas públicas eficazes contra essas atividades ilegais.

Múltiplos fatores estão associados ao número de crimes de homicídios em grandes centros. Segundo Kamaluddin *et al*. (2015), fatores psicológicos dos indivíduos são as variáveis determinantes para a ocorrência de tais atos ilegais. Os autores Short *et al*. (2008) e Alves *et al*. (2015) evidenciaram em seus estudos que existe um padrão espacial que determina a ocorrência de crimes. Por outro lado, Becker (1968), Ehrlich (1973), Wilson e Kelling (1982) e Glaeser, Sacerdote e Scheinkman (1996) verificaram que são os indicadores sociais e econômicos que são os fatores relevantes.

Nota-se que, dependendo da perspectiva e da metodologia adotada pelos pesquisadores, há grandes controvérsias na literatura sobre determinantes de crimes. Algumas explicações para essas diferenças podem ser esclarecidas pelos problemas na seleção de variáveis (LEVITT, 2001), erros relacionados ao levantamento de dados (MALTZ e TARGONSKI, 2002), hipóteses estatísticas não verificadas (GORDON, 2010) e transformações nas variáveis (ALVES *et al*., 2018).

Uma larga escala dos estudos relacionados a análise de determinantes do crime utiliza métodos lineares, tal como a regressão via mínimos quadrados ordinários (ALVES *et al*., 2018). Esses modelos padrão assumem que os preditores possuem exogeneidade fraca, linearidade, variância constante, normalidade dos resíduos e ausência de multicolinearidade. No entanto, dentro das ciências sociais, muitas dessas suposições não são válidas (ALVES *et al*., 2018). E, quando não são válidas ou não verificadas pelos pesquisadores, as conclusões sobre os fatores podem ser potencialmente equivocadas.

Em contrapartida, os modelos de regressão não paramétricos da classe de aprendizado de máquina não fazem as suposições supracitadas. Assim, essas técnicas não paramétricas não possuem suposições sobre o relacionamento das covariáveis com a variável resposta, ou seja, são ferramentas que não fazem pressuposições sobre as distribuições das variáveis e não tem hipóteses sobre a forma funcional entre as mesmas. Desta forma, são técnicas robustas que trazem como principal vantagem a sua grande capacidade de predição, aumentando a acurácia dos modelos preditivos e seu potencial em descobrir possíveis determinantes (BREIMAN, 2001).

Enfim, o presente estudo teve como objetivo identificar os determinantes dos crimes de homicídios no Brasil. Em específico, este trabalho tem dois principais objetivos: 1) utilizar técnicas de aprendizado de máquina para predizer crimes de homicídios nos 5.570 municípios brasileiros, onde os métodos são: Árvore de Regressão, Florestas Aleatórias, *Boosting*, *K-Nearest Neighbors* e Regressão Linear via Mínimos Quadrados e 2) além da busca pelo melhor algoritmo para predição, determinar quais variáveis que mais interferem na ocorrência de crimes de homicídios em nível nacional e fazer uma análise das mesmas.

Considera-se como principais contribuições deste trabalho: i) encontrar determinantes à nível nacional é um grande desafio dentro da ciência quantitativa, pois os métodos tradicionais apresentam muitas quebras de pressuposições devido a heterogeneidade entre os municípios e regiões; ii) a não transformação das variáveis originais, representando um ganho na interpretação dos resultados; iii) além de evidenciar potenciais determinantes para o nível de crime, fazer predições; iv) comparação entre técnicas inovadoras na saúde pública e v) uso de técnicas mais flexíveis do que modelos lineares.

2 Métodos

A variável resposta neste trabalho é o número de crimes de homicídios em cada cidade brasileira (CID-10 X85-Y09) no ano de 2015, e foram utilizadas 31 covariáveis das categorias renda, educação e demográficas, obtidas por meio do IPEADATA e IBGE (Tabela 1 em anexo).

As covariáveis são do ano de 2000 e serão utilizadas para a predição do número de crimes de homicídios 15 anos depois. Este intervalo de tempo foi escolhido pois as características demográficas e socioeconômicas levam um tempo para manifestarem seus efeitos na sociedade, ou seja, considerou-se tal período onde os nascidos em 2000 estariam, conforme sugere a literatura econômica, na idade propensa ao crime (BETTENCOURT *et al*., 2010; ALVES *et al*., 2018).

2.1 Modelos

Árvore de Regressão: A metodologia árvore de regressão é uma técnica não paramétrica que consiste na construção de particionamento recursivo no espaço das covariáveis. Assim, uma árvore cria uma partição do espaço das covariáveis em regiões distintas e disjuntas: $R_1, .., R_j$. A predição para a resposta Y de uma observação com covariáveis **X** que estão em R_k é então dada por $g(\boldsymbol{x}) = \frac{1}{|\{i:\, x_i \in R_k\}|} \sum_{i:\, x_i \in R_k} y_i$. A criação da estrutura da árvore se dividi em duas etapas: I) a criação de uma árvore grande e complexa e II) a poda, para evitar o super-ajuste (IZBICKI e SANTOS, 2018).

O procedimento da poda tem por objetivo tornar a árvore de regressão menor e menos complexa, com o intuito de diminuir a variância do estimador. Para tal, cada nó é retirado, um por vez, e é analisado como o erro de predição varia no conjunto de validação. Assim, decide-se quais os nós que irão permanecer na árvore. Como supracitado, este procedimento garante que, ao diminuir a variância do estimador, corroboramos para evitar o super-ajuste do modelo, ou seja, quando o modelo se adequa demasiadamente as características particulares do conjunto de treinamento e acaba prejudicando seu potencial de generalização para novos dados (BREIMAN, 2001).

Florestas Aleatórias: A técnica de árvore de regressão traz resultados bem interpretáveis, porém, seu poder preditivo é baixo em muitas vezes. A técnica chamada de Florestas Aleatórias visa suprir essa limitação, na qual a principal ideia é a construção de *B* árvores distintas e combinar seus resultados. A criação das florestas aleatórias inicia-se com *B* amostras *boostrap* da amostra original, onde em cada uma cria-se uma árvore diferente e combina seus resultados, assim melhorando, na maioria das vezes, o poder preditivo. Logo, a função de predição consiste em $g(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B} g_b(\boldsymbol{x})$, onde *B* é um *tuning parameter*.

***Boosting*:** Da mesma forma que em florestas aleatórias, o *boosting* consiste na agregação de diferentes estimadores da função de regressão, porém realizada de uma maneira diferente. Formalmente, o algoritmo do *boosting* é: a) definimos g(**x**) = 0 e $r_i = y_i$ para todo *i*; b) para *b* = 1,...,B: i) ajustamos uma árvore com *d* folhas para toda a amostra. Seja $g^b(\boldsymbol{x})$ sua respectiva função de predição, ii) atualizamos g e os resíduos:

$g(\boldsymbol{x}) \leftarrow g(\boldsymbol{x}) + \lambda g^b(\boldsymbol{x})$ e $r_i \leftarrow Y_i - g(x)$. c) retornamos o modelo final $g(\boldsymbol{x})$. Nota-se que os *tuning parameters* aqui são B, d *e* λ. Neste trabalho os estimadores para o método *boosting* foram baseados em árvores.

K-Nearest Neighbors (K-NN): O método do k-vizinhos mais próximos consiste na criação da função de regressão $g(\boldsymbol{x})$ para uma dada configuração das covariávies **X** com base nas respostas Y dos k-vizinhos mais próximos a **X**. Formalmente, temos $g(\boldsymbol{x}) = \frac{1}{k}\sum_{i \in N_x} y_i$, em que N_x é o conjunto das k observações mais próximas de **X**, *i.e*, $N_x = \{i \in \{i, \dots, n\} : d(x_i, x) \leq d_x^k\}$, e d_x^ké a distância do k-ésimo vizinho mais próximo de **X** a **X**. Nota-se que o *tunning parameter* é o *k*, e o mesmo pode ser escolhido via validação cruzada, que será discutido na próxima subseção. Salientando que, um valor alto de *k* leva a um modelo simples (viés alto com uma variância baixa). Por outro lado, um valor baixo de *k* acarreta em um estimador com variância alta, mas um viés baixo.

2.2 Medidas de qualidade de ajuste e extração dos determinantes

Seguiremos a cultura chamada de *algorithmic modeling culture*, onde é dominada pela comunidade de aprendizado de máquina (BREIMAN, 2001). Nesta cultura, o principal objetivo é a predição de novas observações e conhecer seus principais determinantes, e não inferenciais (IZBICKI e SANTOS, 2018).

Para controlar o super-ajuste, sub-ajuste, e *o trade-off* entre viés e variância, os modelos e os parâmetros foram escolhidos por validação cruzada. Além disso, foi utilizado 80% das observações para treinamento (4.456 municípios) e 20% para validação (1.114 municípios). Para avaliar a qualidade preditiva dos modelos, utilizou-se as seguintes métricas: *Relative Absolute Error* (RAE), *Root Mean Squared Error* (RMSE) e a Variância Explicada.

O método de validação cruzada *k-fold* consiste em repartir o conjunto de treinamento original em *k*-subconjuntos. Para cada subconjunto, estimamos o método sem a presença da *k*-1 parte e verificamos o erro médio no conjunto não utilizado durante o treino. A estimativa do erro de predição é dada pela média dos erros nos *k* subconjuntos. A metodologia 10-fold foi selecionada neste trabalho, pois caso tivéssemos escolhido *k* igual a *N* (*leave-one-out*), teríamos um alto custo computacional devido ao ajuste do modelo ser realizado *N* vezes. Diminuindo o valor de *k*, diminuiremos a variância, mas podemos aumentar o viés do estimador. Na prática, costuma-se utilizar *k* variando de 5 a 10 (KOHAVI *et al*; 1995; BREIMAN, 1996).

Portanto, o presente estudo tem o seguinte fluxograma: i) em um primeiro momento, selecionar o melhor modelo preditivo, pois assim teremos o modelo com o maior número de acertos quanto ao número de crimes de homicídios e, ii) com o melhor modelo em mãos, verificar e analisar quais são os principais determinantes.

Nota-se que, na fase (i) iremos realizar perturbações na amostra de treinamento, onde iremos obter 1000 amostras diferentes, com o objetivo de não ter viés na escolha da amostra de treinamento e teste. Para tanto, utilizamos a função *sample* do software R 3.4.1, que tem por objetivo realizar amostras aleatórias onde não utilizamos a reposição e todos os elementos possuem probabilidades iguais de pertencer ao grupo sorteado.

A extração da importância dos determinantes (ii) baseia-se em metodologias ligeiramente diferentes entre as técnicas abordadas. Para a árvore de regressão, as variáveis com maiores pesos são as aquelas nas quais a combinação leva a uma partição

com predições de menor erro quadrático em todos os nós. Assim, o tamanho de cada ramo na árvore gerada é proporcional à diminuição do erro quadrático médio que ocorreu quando a respectiva partição foi criada. Logo, ramos grandes indicam uma importância maior da covariável na predição da variável resposta.

Já os métodos Florestas Aleatórias e *Boosting* baseiam-se na média de quanto cada variável foi importante em cada árvore, ou seja, o quanto cada variável contribuiu para diminuir o erro quadrático médio. Logo, para os três métodos supracitados as importâncias são um subproduto da construção matemática dos próprios métodos. Já o método K-NN, por exemplo, pode se utilizar o método de análise de sensibilidade de King e Perera (2013) ou Yang *et al*. (2017).

3 Resultados

Como descrito na seção anterior, esse primeiro resultado tem por objetivo obter um algoritmo com melhor poder preditivo utilizando as variáveis independentes propostas e as métricas de seleção de modelos adotadas e, em seguida, analisar quais foram os principais determinantes de acordo com o melhor modelo preditivo. A Tabela (1) em anexo apresenta as variáveis, interpretações e suas médias, medianas e desvios-padrão.

De acordo com a Figura (1) nota-se que as variáveis independentes selecionadas neste estudo trazem diferentes magnitudes de correlações lineares, onde se verifica que muitas possuem associações fortes e positivas (>0,7) e também negativas (<-0,7). É evidente que em métodos tradicionais de regressão lineares isso acarretaria em problemas inferenciais devido a multicolinearidade, porém, os métodos propostos neste estudo trabalham com essa natureza.

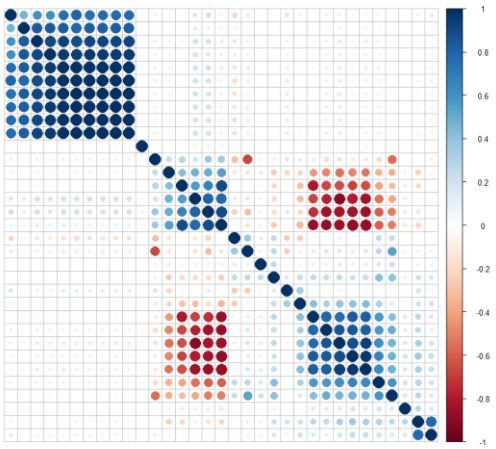

Figura 1 – Correlação entre as covariáveis deste estudo.

Para fazer as predições considerando o 10-*fold*, obtivemos os seguintes *tuning parameters*: Florestas Aleatórias: $B = 500$; *Boosting*: $B = 1000$, $d = 3$, $\lambda = 0.02$; e K-NN: $k = 15$. De acordo com a Tabela (2), as métricas de qualidade de ajustes preservaram a ordem dos modelos, ou seja, notamos que as métricas concordam que o melhor método

para realizar a predição foi o Florestas Aleatórias, onde o mesmo explica, em média, 82% da variabilidade dos crimes de homicídios em nível nacional. Em ordem decrescente vem o *Boosting* (78%), Árvore de Regressão (67%), Regressão via Mínimos Quadrados (61%) e por último o K-NN (39%).

Tabela 2 – Métricas de qualidade de ajuste dos modelos

Modelo	MAE (s.d)	RMSE (s.d)	RAE (s.d)	Variância Explicada (s.d)
Árvore	7,279 (0,890)	1054,3 (605,6)	0,510 (0,063)	67,10 (16,0)
Florestas	5,104 (0,644)	608,5 (450,1)	0,357 (0,039)	81,8 (8,4)
Boosting	5,486 (0,723)	717,6 (462,2)	0,383 (0,040)	78,1 (10,0)
K-NN	8,740 (1,238)	2056,2 (1182,8)	0,609 (0,052)	39,6 (12,7)
Regressão	8,447 (0,688)	1420,1 (1415,8)	0,596 (0,084)	61,7 (26,6)

Porém, para verificarmos se a diferença entre o método Florestas Aleatórias e os demais é estatisticamente significante, realizamos o teste de hipótese ANOVA e o teste *t-student* pareado para comparação entre pares para as métricas. Devido as métricas MAE, RMSE e RAE terem mantido a mesma ordem de escolha do melhor para pior modelo, será exposto neste texto somente os resultados para o MAE.

De acordo com as Tabela (3) e (4) notamos que os valores-p estão abaixo do nível de significância adotado de 5%. Portanto, rejeitamos a hipótese nula, ou seja, pelo menos um dos modelos apresenta um padrão médio das métricas distinto dos demais, tanto para o MAE quanto para a variância explicada.

Tabela 3 – Teste ANOVA para a métrica MAE

	GL	Soma de Quadrados	Média da Soma de Quadrados	Valor F	Pr(>F)
Modelos	4	11081	2770,3	3704	<2e-16 ***
Resíduo	4995	3736	0,7		

Tabela 4 - Teste ANOVA para a métrica variância explicada

	GL	Soma de Quadrados	Média da Soma de Quadrados	Valor F	Pr(>F)
Modelos	4	111	27,76	1065	<2e-16 ***
Resíduo	4995	130,2	0,02		

As Tabelas (5) e (6) apresentam os valores-p para o teste *t-student* pareado considerando as mesmas medidas, ou seja, MAE e variância explicada.

Tabela 5 – Valor-p do teste *t-student* pareado para o MAE

	Árvore	Boosting	KNN	Regressão
Boosting	<2e-16	-	-	-
KNN	<2e-16	<2e-16	-	-
Regressão	<2e-16	<2e-16	<2e-16	-
RF	<2e-16	<2e-16	<2e-16	<2e-16

Tabela 6 – Valor-p do Teste *t-student* pareado para a variância explicada

	Árvore	Boosting	KNN	Regressão
Boosting	<2e-16	-	-	-
KNN	<2e-16	<2e-16	-	-
Regressão	2,2e-07	<2e-16	<2e-16	-
RF	<2e-16	<2e-16	<2e-16	<2e-16

Com base nos resultados das Tabelas (5) e (6), podemos concluir que o modelo de Florestas Aleatórias é significativamente distinto dos demais modelos, pois os valores apresentados na Tabela (2) são maiores do que para os outros modelos e os valores-p são menores do que 0,05 nas Tabelas (5) e (6).

A Figura (2) apresenta os valores preditos e observados para uma amostra em específico, onde na parte superior temos em escala real dos dados e, para uma melhor visualização, em escala log na parte inferior. Essa figura corrobora com o bom ajuste dos modelos Florestas Aleatórias e *Boosting*, em que notamos a maior concentração dos dados ao redor da reta de 45º entre os preditos e observados. Atinamos pelo método de Florestas Aleatórias a grande acurácia do método mesmo quando o número de crimes de homicídios ocorre com uma frequência muito maior do que as demais cidades, típico acontecimento em grandes centros urbanos, tais como Recife e São Paulo, por exemplo.

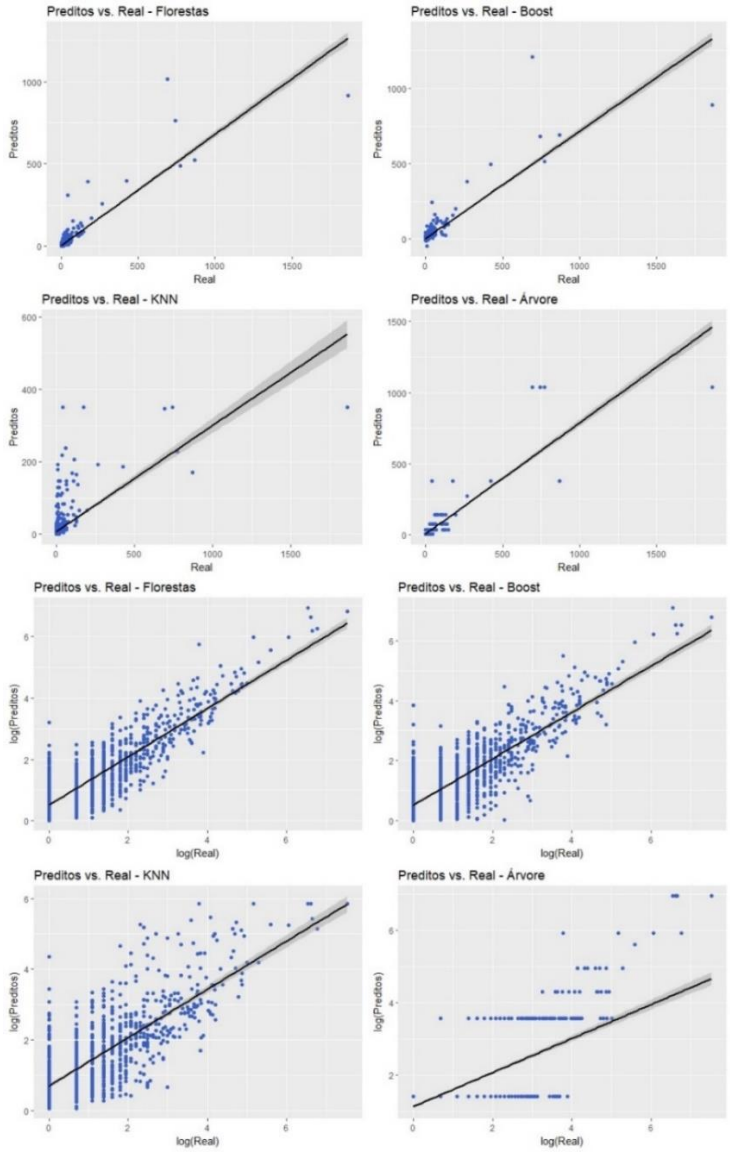

Figura 2 – Valores Preditos e observados em escala real (4 primeiros) e em escala log (4 últimos).

Tabela 7 -Importância de cada Fator no modelo de Florestas Aleatórias

Fatores	Importância (Média e Desvio-Padrão)	
População Jovem	8,99%	2,93%
Saneamento Básico	8,28%	2,27%
População Total	8,10%	2,03%
População Economicamente Ativa	7,36%	1,88%
População Urbana	7,15%	1,44%
PIB	6,58%	0,45%
Mulher Chefe da Família	5,60%	0,20%
Pessoas Pobres entre 0 a 14 anos	5,30%	0,07%
Proporção de Pessoas que ganham até meio Salário Mínimo	5,23%	0,13%
População Masculina	4,16%	0,16%
Índice de Desigualdade de *Theil*	3,90%	0,04%
Renda Média Per Capita	2,73%	0,03%
Despesas com Segurança	2,68%	0,33%
Anos de Estudo Médio - Pessoas com 25 anos ou mais	2,35%	0,35%
Despesas com Cultura e Educação	2,35%	0,37%
Razão entre Renda dos Pobres e Ricos	2,33%	0,38%
Mulher entre 10 a 14 anos com Filhos	2,24%	0,45%
Mulher entre 15 a 17 anos com Filhos	2,21%	0,46%
Evasão Escolar entre 7 a 14 anos	1,64%	0,50%
Mortalidade Infantil	1,61%	0,55%
Pessoas Indigentes entre 0 a 14 anos	1,50%	0,57%
Taxa de Fecundidade	1,36%	0,59%
Probabilidade de Sobrevivência até os 40 anos	1,26%	0,61%
IDHM	1,04%	0,62%
Índice de *Gini*	1,01%	0,63%
Taxa de Desemprego pessoas 16 anos ou mais	0,70%	0,63%
Taxa de Analfabetismo	0,67%	0,64%
Frequência Escolar de pessoas entre 7 a 22 anos	0,67%	0,68%
Despesas com Desenvolvimento Regional	0,50%	0,83%
Taxa de Participação das Mulheres	0,27%	0,60%
Evasão Escolar entre 15 a 17 anos	0,24%	0,30%

Notamos que, o modelo de florestas aleatórias obteve a maior acurácia na predição e, por esse motivo esta técnica será utilizada para conhecermos quais variáveis são importantes para conhecer o fato gerador do comportamento criminoso. Devido às perturbações realizadas na escolha dos nossos dados para treinamento e para teste, as variáveis importantes podem mudar de acordo com os dados aleatoriamente escolhidos.

Para controlar esse efeito, a Tabela (7) apresenta o ranking calculado verificando a média da importância de cada variável em cada amostra adotada (média e desvio-padrão entre as amostras). A importância de cada fator no modelo foi calculada pela sua capacidade em diminuir seu erro médio na predição do algoritmo, portanto, quanto maior a porcentagem, mais influente é o fator.

Por fim, adotou-se um corte de pelo menos 5% de importância no modelo de Florestas Aleatórias para apresentar a interpretação de cada fator seguindo a literatura econômica do crime, apresentada na próxima seção.

4 Discussão e considerações finais

Devido as significativas transformações no espaço urbano que estão relacionadas ao desenvolvimento econômico, as cidades cresceram como consequência das ofertas de empregos nos setores industriais, de serviços e de transporte. Uma das principais consequências é um crescimento desordenado e um controle ineficiente das autoridades responsáveis, causando um aumento nos números de criminalidade não apenas nos grandes centros, mas também passou a ser um problema de cidades menores.

Desta forma, o presente estudo investigou metodologias na predição e determinantes dos crimes de homicídios no Brasil. Em uma primeira etapa, verificamos que os resultados indicaram que o modelo de florestas aleatórias obteve boa performance nas métricas de avaliação de predição e evidenciou que o crime é ligeiramente dependente de fatores sociais, econômicos e demográficos, onde o mesmo explicou, em média, 82% da variabilidade dos crimes de homicídios em nível nacional.

Quanto aos determinantes, notamos que entre as 31 variáveis independentes utilizadas, a população jovem foi a que mais teve efeito no melhor modelo preditivo. De uma perspectiva econômica, os jovens recebem menores salários, por isso teriam um custo de oportunidade menor ao exercer atividades criminosas (VALLE e MARZANO, 2011). E, além disso, há a razão do auto consentimento de serem menos propensos a controles sociais. O estudo de Zaluar (1994) mostrou que, uma pesquisa realizada com jovens, a via criminosa era uma saída para obtenção de renda para a entrada na chamada sociedade de consumo.

O segundo determinante mais presente nos modelos de florestas aleatórias foi o saneamento básico. Esse resultado nos leva a interpretação de uma decorrência do rápido processo de urbanização da sociedade brasileira, destacando-se que o crime organizado encontrou espaço para o crescimento no país (PROCÓPIO e TOYOSHIMA, 2014). Os autores Glaeser, Sacerdote e Scheinkman (1996) mostraram que em regiões mais urbanizadas é comum que haja uma maior troca de informações entre os criminosos, o que acarretaria em menores custos de planejamento e execução do ato ilícito. Outra razão para este relacionamento é a qualidade de vida dos mais pobres sendo comprometida pela falta de acesso a serviços públicos básicos, como o saneamento. E, como supracitado, a prática

criminosa, surge como uma possibilidade de obtenção de recursos para uma melhora de vida dessa população.

A terceira e quinta variável com maiores impactos foram a população total e a população urbana. Esses fatores funcionam como um controle demográfico, assim como discutido em Teixeira (2011). Conforme o autor, a variável população urbana tem-se como hipóteses que, em conglomerados o efeito punição de Becker (1968) seja percebido pelos criminosos como pouco eficaz, ou seja, para o indivíduo o custo da prática delituosa é menor do que os benefícios esperados, dado que através do conglomerado sua probabilidade de ser pego tende a diminuir, e a interação de criminosos com futuros delinquentes ocorra de maneira mais intensa. De forma correlata, segundo Ehrlich (1973), o tamanho da população e sua densidade são negativamente relacionados à probabilidade de punição, pois em uma região mais densa, um agente criminoso consegue fugir mais facilmente dos órgãos responsáveis pela defesa do local.

O quarto fator predominante no melhor modelo foi a população economicamente ativa. Embora não percebida diretamente, essa variável explica espontaneamente os crimes, no qual mostra o relacionamento positivo entre eles, ou seja, as práticas de atos ilegais concentram-se em regiões de negócios e na maior concentração de população economicamente ativa (COHEN *et al*., 1981).

A sexta variável é o PIB, tendo a mesma uma interpretação ambígua na literatura. Por exemplo, os autores Moreira et al. (2017) explicaram que uma riqueza maior implica em uma população com maior poder aquisitivo, o que permite acesso a uma cesta de bens com itens de segurança e defesa, inibindo a atividade criminosa. Por outro lado, nosso resultado corrobora com os de Cohen et al. (1981), em que os autores argumentaram que uma maior riqueza implica em potenciais vítimas economicamente mais atrativas, acarretando em uma contribuição para elevar o retorno da atividade criminosa.

O sétimo fator predominante foi a variável famílias chefiadas por mulheres. De acordo com Santos e Kassouf (2007), esta variável está relacionada com o grau de desorganização social, no qual a mesma tem um efeito positivo sobre o nível de criminalidade. Fajnzylber *et al*. (2002) obtiveram resultados semelhantes, reforçando que a desagregação familiar de fato interfere na decisão do indivíduo em cometer ou não crimes. Os criminologistas, em geral, associam crime, instabilidade familiar e distúrbios emocionais sofridos pelos indivíduos durante o período de infância e adolescência (KELLY, 2000). Esse resultado é justificado nessas situações, pois, em ambientes familiares conturbados, menores serão os custos morais do indivíduo em relação ao ato criminoso.

Em sequência, a próxima variável foi a proporção de pessoas entre 0 a 14 anos que são pobres. Zaluar (1994) relata que, tanto nos relatos das experiências de vida quanto nas respostas recebidas em seus estudos, os criminosos entrevistados referiram-se sempre a uma fase crucial da vida, que começa em torno dos 14 anos, como um ponto inicial no envolvimento com a criminalidade. A junção da proporção de pessoas pobres entre 0 a 14 anos e instabilidade familiar podem ser fatores determinantes para o aliciamento dos jovens ao crime. Conforme tal autora, o aliciamento dos jovens na substituição de outros bandidos vai a favor da lógica do jogo de mercado do sistema capitalista de produção: “maior será o lucro quanto mais barata for a mão-de-obra empregada.

A última e nona variável a ser analisada é a proporção da população que ganha meio salário mínimo ou menos. A suposição básica e já discutida relacionada a variável renda é que os indivíduos respondem a incentivos econômicos. Portanto, salários mais baixos

implicam menores oportunidades no mercado de trabalho (GOULD *et al*., 2002), o que diminui o custo alternativo de atividades ilegais e, consequentemente, estimula a entrada e ou a permanência no crime.

Os achados deste estudo indicaram que a aplicação de métodos de aprendizado de máquina são boas ferramentas para verificar e analisar quais variáveis impactam os níveis de crimes. Sendo assim, os resultados evidenciados neste trabalho são possíveis ferramentas para gestores de políticas públicas, onde os mesmos podem realizar estratégias de médio e longo prazo na redução das taxas de criminalidade.

Como trabalhos futuros, consideramos importantes os seguintes aspectos: (i) neste trabalho foi abordado os crimes de homicídios de acordo com a classificação da OMS CID-10, porém, há espaços para estudar se aspectos econômicos estariam relacionados mais fortemente com crimes de patrimônios e se fatores sociais com crimes de homicídios; (ii) a inclusão de outras variáveis, tais como *proxies* para eficiência policial e (iii) nossos achados vão de encontro com resultados da literatura, mostrando a relação entre o tamanho da população e o número de homicídios, porém uma questão importante a ser tratada posteriormente é fazer a mesma análise separando a amostra em relação ao tamanho das cidades para evidenciar se os fatores permanecem o mesmo (BANERJEE *et al*., 2015). Em que, a modificação na importância desses fatores pode levar a formulação de diferentes políticas públicas, que poderiam ser diferenciadas para cidades de pequeno porte e cidades de médio/grande porte.

Agradecimentos

Os autores agradecem aos revisores e editores pelas sugestões.

LOPES, L. P.; FELIX, S. V. Determinants and prediction of homicide crimes in Brasil: a machine learning approach. *Rev. Bras. Biom*. Lavras, v.37, n.2, p.272-289, 2019.

- *ABSTRACT: Throughout history, organized societies have attempted to prevent crime by various approaches, and the social justification for understanding what features associated with crime is their use to achieve effective public policies against such illegal activities. In this context, the objective of this work is to identify the economic, social and demographic determinants of homicide crimes in Brazil. And, as a secondary objective, the prediction of the level of crimes in national territory. As a methodology, this is a quantitative study where Regression Tree, Random Forest, Boosting and K-Nearest Neighbors (K-NN) methods were used as alternative tools to traditional linear models, such as regression via ordinary least squares. The data analyzed indicated that among the 31 covariates used, 9 had the greatest impacts on violence at the national level, such as the size of the young population, basic sanitation, total population size, economically active population, urban population, GDP , female heads in the family, poor people between 0 and 14 years old and the proportion of people earning up to half a minimum wage, where each factor was discussed according to the economic literature of the crime. In addition, the Random Forest model explained, on average, 82% of the variability of homicide crimes at the national level. We believe that this approach helps to produce more robust responses on the effects of social, economic, and demographic factors on crime and is therefore a new tool for guiding policymakers.*

- *KEYWORDS: Random forests; public policy; violence; socioeconomic factors.*

Referências

ALVES, L. G. A.; LENZI, E. K.; MENDES R, S.; RIBEIRO, H. V. Spatial correlations, clustering and percolation-like transitions in homicide crimes. *Europhysics Letters*, 2015.

ALVES, L. G. A.; RIBEIRO, H. V.; RODRIGUES, F. A. Crime prediction through urban metrics and statistical learning**.** *Physica A: Statistical Mechanics and its Applications*, v.505, p.435-443, 2018.

BANERJEE, S.; VAN HENTENRYCK, P.; CEBRIAN, M. Competitive dynamics between criminals and law enforcement explains the super-linear scaling of crime in cities. *Palgrave communications*, v.1, p.15022, 2015.

BECKER, G. S. Crime and punishment: An economic approach. *The Journal of Political Economy*, v.76, n.2, p.169-217, 1968.

BETTENCOURT, L. M.; LOBO, J.; STRUMSKY, D.; WEST, G, B. Urban scaling and its deviations: Revealing the structure of wealth, innovation and crime across cities, *PLoS ONE*, 2010.

BREIMAN, L. Bagging predictors. *Machine learning*, v.24, n.2, p.123-140, 1996.

BREIMAN L. Random forests. *Machine Learning*, v.45, n.1, p.5–32, 2001.

COHEN, L. E.; KLUEGEL, J. R.; LAND, K. C. Social inequality and predatory criminal victimization: an exposition and test of a formal theory. *American Sociological Review*, v.46, n.5, p.505-524, 1981.

DATAGOVIN - *Open government data platform India*, 2018. Disponível em: <https://data.gov.in/dataset-group-name/crime-statistics>. Acesso em: 21 nov. 2018.

EHRLICH, I. Participation in Illegitimate Activities: a theoretical and empirical investigation. *The Journal of Political Economy*, v.81, n.3, p.521-565, 1973.

EUROSTAT - *Your key to European statistics*, 2018. Disponível em: <https://ec.europa.eu/eurostat/web/crime/database>. Acesso em: 21 nov. 2018.

FAJNZYLBER, P.; LEDERMAN, D.; LOYAZA, N. Inequality and violent crime. *Journal of Law and Economics*, v.45, n.1, p.1-40, 2002.

GLAESER, E.; SACERDOTE, B.; SCHEINKMAN, J. Crime and social interactions. *Quarterly Journal of Economics*, v.111, n.2, 1996.

GORDON, M. B.; IGLESIAS, J. R.; SEMESHENKO, V.; NADAL, J. P. Crime and punishment: the economic burden of impunity, *The European Physical Journal* B, v.68, 2009.

GORDON, M. B. A random walk in the literature on criminality: A partial and critical view on some statistical analyses and modelling approaches, *European Journal of Applied Mathematics*, v.21, 2010.

GOULD, E. D.; WEINBERG, B. A.; MUSTARD, D. B. Crime rates and Labor Market Opportunities in The United States: 1979 – 1997. *The Review of Economics and Statistics*, v.84, n.1, p.45-61, 2002.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. *Pesquisa Nacional de Saúde 2013*. Rio de Janeiro: IBGE, 2013. Disponível em: <https://ww2.ibge.gov.br/home/estatistica/populacao/pns/2013_vol2/default.shtm>. Acesso em: 21 nov. 2018.

IPEA - INSTITUTO DE PESQUISA ECONÔMICA APLICADA. *Atlas da Violência*. Brasília, 2018. Disponível em: <http://www.ipea.gov.br/atlasviolencia/>. Acesso em: 21 nov. 2018.

IZBICKI, R; SANTOS, T. M. *Machine learning sob a ótica estatística*: Uma abordagem preditivista para estatística com exemplos em R. Notes, 2018.

KAMALUDDIN, M. R.; SHARI, N.; OTHMAN, A.; Ismail, K. H.; SAAT, G. A. M. Linking psychological traits with criminal behaviour: A review. *Journal of Psychiatry*, 2015.

KELLY, M. Inequality and crime. T*he Review of Economics and Statistics*, v.82, n.4, p.530–539, 2000.

KING, D. M.; PERERA, B. J. C. Morris method of sensitivity analysis applied to assess the importance of input variables on urban water supply yield–a case study. *Journal of hydrology*, v.477, p.17-32, 2013.

KOHAVI, R. *et al*. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Ijcai*. 1995. p.1137-1145.

LEVITT, S. D. Alternative strategies for identifying the link between unemployment and crime, *Journal of Quantitative Criminology*, 2001.

MALTZ, M. D.; TARGONSKI, J. A note on the use of county-level ucr data, *Journal of Quantitative Criminology*, 2002.

MOREIRA, C. G.; KASSOUF, A. L.; JUSTUS, M. Crimes Patrimoniais não Registrados: Mensuração e Determinantes. *Anais do Encontro Nacional de Economia*, 2017.

PROCÓPIO, D. P.; TOYOSHIMA, S. H. Fatores associados à criminalidade violenta no Brasil. *Análise Econômica* 35, 2014.

R CORE TEAM. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/, 2018.

SANTOS, M.; KASSOUF, A. Uma investigação econômica da influência do mercado de drogas ilícitas sobre a criminalidade brasileira. *Economia*, v.8, n.2, p.187-210, 2007.

SHORT, M. B.; D'ORSOGNA, M. R.; PASOUR, V. B.; TITA. G. E.; BRANTINGHAM, P. J.; BERTOZZI, A. L.; CHAYES, L. B.; SPELMAN, W. Specifying the relationship between crime and prisons, *Journal of Quantitative Criminology*, 2008.

TEIXEIRA, E. C. *Dois ensaios acerca da relação entre criminalidade e educação*. Tese em Economia Aplicada. Escola Superior de Agricultura Luiz de Queiroz - ESALQ/USP. 102 p. 2011.

UCR - Uniform rrime reporting. USA, 2018. Disponível em: <https://ucr.fbi.gov/crime-in-the-u.s>. Acesso em: 21 nov. 2018.

VALLE, P. A.; MARZANO, V. *Economia e criminalidade: uma Análise das Mesorregiões de Minas Gerais no Período 2005-2007*. Curso de Ciências Econômicas da Universidade Federal de Goiás -FACE, 2011.

WILSON, J. Q.; KELLING, G. L. *Broken windows*, Atlantic Monthly, 1982.

YANG, H. *et al*. Sobol sensitivity analysis for governing variables in design of a plate-fin heat exchanger with serrated fins. *International Journal of Heat and Mass Transfer*, v.115, p.871-881, 2017.

ZALUAR, A. *Condomínio do diabo*. Rio de Janeiro: UFRJ Editora, 1994.

Recebido em 27.07.2018

Aprovado após revisão em 19.12.2018

ANEXO

Tabela 1 - Interpretação e algumas medidas resumos das covariáveis

Variável	Interpretação/Fonte	Mediana	Média	Desvio
Homicídios	Número de crimes de homicídios em cada cidade brasileira (CID-10 X85-Y09)	1,00	10,54	60,17
População Masculina	Porcentagem da população masculina – IBGE	0,50	0,50	0,01
População Total	População Total em cada cidade – IBGE	10417	30826	187278,5
População Urbana	População vivendo na cidade - IBGE	5298	24121	178806,5
População Jovem	População de 15 a 24 anos em cada cidade - IBGE	2311	6406	34151,08
Taxa de Analfabetismo	Taxa de Analfabetismo das pessoas de 15 anos ou mais – IPEADATA	17,20	20,78	12,19
Renda Média Per Capita	*Renda* domiciliar *per capita* - *IPEADATA*	305,53	336,98	199,89
Índice de Gini	O Coeficiente de Gini é uma medida de desigualdade - IPEADATA	0,55	0,55	0,06
Proporção da População que ganha até Meio Salário	Porcentagem da População que ganha até Meio Salário - IBGE	65,59	64,11	20,64
Taxa de Desemprego	Se refere à desocupação oficial no país - IBGE	9,68	10,32	5,93
População Economicamente Ativa	É um conceito elaborado para designar a população que está inserida no mercado de trabalho ou que, de certa forma, está procurando se inserir nele para exercer algum tipo de atividade remunerada - IBGE	4111	13485	91393,18
PIB	Produto Interno Bruto - IPEADATA	27401	214456	2601045
Abastecimento de Água	População com abastecimento de Água - IPEADATA	10320	30567	185728,4
Probabilidade de Sobrevivência até 40 anos	IPEADATA	90,67	89,28	5,85

Tabela 1 (Continuação) - Interpretação e algumas medidas resumos das covariáveis

Variável	Interpretação/Fonte	Mediana	Média	Desvio
Taxa de Fecundidade	É uma estimativa do número médio de filhos que uma mulher teria até o fim de seu período reprodutivo - IBGE	2,68	2,87	0,75
Razão da Renda entre Rico e Pobre	Ée a razão da renda entre os 20% mais ricos e os 40% mais pobres - IPEADATA	18,00	24,62	57,79
Índice de Desigualdade de Theil	É uma medida estatística da distribuição de renda - IPEADATA	0,51	0,52	0,11
Mulheres entre 10 a 14 anos com Filhos	Mulheres com Filhos nascidos - IBGE	0,20	0,44	0,62
Mulheres entre 15 a 17 anos com Filhos	Mulheres com Filhos nascidos - IBGE	8,03	8,72	4,55
Mortalidade Infantil	Taxa de Mortalidade Infantil - IBGE	30,38	34,56	18,54
Despesas com Segurança	Total em milhões gastos com Despesas de Segurança - IBGE	0,00	66651	1537076
Despesas com Desenvolvimento Regional	Total em milhões gastos com Despesas de Desenvolvimento Regional - IBGE	0,00	6050	85806,14
Despesas com Educação	Total em milhões gastos com Despesas de Educação - IBGE	1415000	3931000	2540743
Média de anos de estudo – 25 anos ou mais	*Média de anos de estudo* das pessoas de 25 *anos ou mais* de idade - IPEADATA	4,01	3,99	1,28
Evasão escolar – 7 a 14 anos	Porcentagem de Evasão entre 7 a 14 anos - IBGE	5,08	6,17	4,75
Evasão escolar – 15 a 17 anos	Porcentagem de Evasão entre 15 a 17 anos - IBGE	26,88	27,64	8,28
Frequência Escolar – 7 a 22 anos	Porcentagem da Frequência Escolar de 7 a 22 anos - IBGE	78,25	77,81	7,16
IDHM	Índice de Desenvolvimento Humano – IBGE	0,70	0,69	0,08

Tabela 1 (Continuação) - Interpretação e algumas medidas resumos das covariáveis

Variável	Interpretação/Fonte	Mediana	Média	Desvio
Mulheres Chefes na Família	Porcentagem de *famílias* chefiadas por *mulheres* - *IBGE*	4,93	5,03	1,73
Taxa de Participação de Mulheres	T*axa* global de *participação* das *mulheres*- IBGE	0,38	0,38	0,10
Pessoas Indigentes – 0 a 14 anos	Número de miseráveis – Em Milhares - IBGE	30,22	34,01	21,8
Pessoas Pobres – 0 a 14 anos	Pessoas que vivem na linha de pobreza, em Milhares - IBGE	60,20	58,53	23,2