ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Cerqueira, Daniel; Lins, Gabriel de Oliveira Accioly

Working Paper

Mapa dos homicídios ocultos no Brasil entre 1996 e 2021

Texto para Discussão, No. 3015

Provided in Cooperation with:
Institute of Applied Economic Research (ipea), Brasília

Suggested Citation: Cerqueira, Daniel; Lins, Gabriel de Oliveira Accioly (2024) : Mapa dos homicídios ocultos no Brasil entre 1996 e 2021, Texto para Discussão, No. 3015, Instituto de Pesquisa Econômica Aplicada (IPEA), Brasília, https://doi.org/10.38116/td3015-port

This Version is available at:
https://hdl.handle.net/10419/300321

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

CC BY https://creativecommons.org/licenses/by/2.5/br/

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

TEXTO PARA DISCUSSÃO

3015

MAPA DOS HOMICÍDIOS OCULTOS NO BRASIL ENTRE 1996 E 2021

DANIEL RICARDO DE CASTRO CERQUEIRA
GABRIEL DE OLIVEIRA ACCIOLY LINS

ipea

TEXTO PARA DISCUSSÃO

3015
Brasília, junho de 2024

MAPA DOS HOMICÍDIOS OCULTOS NO BRASIL ENTRE 1996 E 2021

DANIEL RICARDO DE CASTRO CERQUEIRA[1]
GABRIEL DE OLIVEIRA ACCIOLY LINS[2]

1. Técnico de planejamento e pesquisa na Diretoria de Estudos e Políticas do Estado, das Instituições e da Democracia do Instituto de Pesquisa Econômica Aplicada (Diest/Ipea). *E-mail*: daniel.cerqueira@ipea.gov.br. Orcid: https://orcid.org/0000-0002-4083-9535.

2. Pesquisador bolsista do Subprograma de Pesquisa para o Desenvolvimento Nacional (PNPD) na Diest/Ipea. *E-mail*: gabriel.lins@ipea.gov.br. Orcid: https://orcid.org/0000-0002-9360-2902.

Governo Federal

Ministério do Planejamento e Orçamento
Ministra Simone Nassar Tebet

ipea Instituto de Pesquisa Econômica Aplicada

Fundação pública vinculada ao Ministério do Planejamento e Orçamento, o Ipea fornece suporte técnico e institucional às ações governamentais – possibilitando a formulação de inúmeras políticas públicas e programas de desenvolvimento brasileiros – e disponibiliza, para a sociedade, pesquisas e estudos realizados por seus técnicos.

Presidenta
LUCIANA MENDES SANTOS SERVO

Diretor de Desenvolvimento Institucional
FERNANDO GAIGER SILVEIRA

Diretora de Estudos e Políticas do Estado, das Instituições e da Democracia
LUSENI MARIA CORDEIRO DE AQUINO

Diretor de Estudos e Políticas Macroeconômicas
CLÁUDIO ROBERTO AMITRANO

Diretor de Estudos e Políticas Regionais, Urbanas e Ambientais
ARISTIDES MONTEIRO NETO

Diretora de Estudos e Políticas Setoriais, de Inovação, Regulação e Infraestrutura
FERNANDA DE NEGRI

Diretor de Estudos e Políticas Sociais
CARLOS HENRIQUE LEITE CORSEUIL

Diretor de Estudos Internacionais
FÁBIO VÉRAS SOARES

Chefe de Gabinete
ALEXANDRE DOS SANTOS CUNHA

Coordenadora-Geral de Imprensa e Comunicação Social
GISELE AMARAL

Ouvidoria: http://www.ipea.gov.br/ouvidoria
URL: http://www.ipea.gov.br

Texto para Discussão

Publicação seriada que divulga resultados de estudos e pesquisas em desenvolvimento pelo Ipea com o objetivo de fomentar o debate e oferecer subsídios à formulação e avaliação de políticas públicas.

© Instituto de Pesquisa Econômica Aplicada – **ipea** 2024

Cerqueira, Daniel Ricardo de Castro
Mapa dos homicídios ocultos no Brasil entre 1996 e 2021 / Daniel Ricardo de Castro Cerqueira, Gabriel de Oliveira Accioly Lins. – Brasília, DF: Ipea, 2024.

Inclui Bibliografia.
ISSN 1415-4765

1. Aprendizado de Máquina. 2. Subnotificação de Homicídio. 3. Sistema de Informação sobre Mortalidade (SIM). 4. Brasil. I. Lins, Gabriel de Oliveira Accioly. II. Instituto de Pesquisa Econômica Aplicada. III. Título.

CDD 364.0981

Ficha catalográfica elaborada por Elizabeth Ferreira da Silva CRB-7/6844.

Como citar:

CERQUEIRA, Daniel Ricardo de Castro; LINS, Gabriel De Oliveira Accioly. **Mapa dos homicídios ocultos no Brasil entre 1996 e 2021**. Brasília, DF: Ipea, jun. 2024. 64 p. : il. (Texto para Discussão, n. 3015). DOI: http://dx.doi.org/10.38116/td3015-port

JEL: C18; C81; K42.
DOI: http://dx.doi.org/10.38116/td3015-port

As publicações do Ipea estão disponíveis para download gratuito nos formatos PDF (todas) e ePUB (livros e periódicos).

Acesse: http://www.ipea.gov.br/portal/publicacoes

As opiniões emitidas nesta publicação são de exclusiva e inteira responsabilidade dos autores, não exprimindo, necessariamente, o ponto de vista do Instituto de Pesquisa Econômica Aplicada ou do Ministério do Planejamento e Orçamento.

É permitida a reprodução deste texto e dos dados nele contidos, desde que citada a fonte. Reproduções para fins comerciais são proibidas.

SUMÁRIO

SINOPSE

ABSTRACT

1 INTRODUÇÃO 6

2 DEFININDO AS TIPOLOGIAS DE ÓBITOS POR INTENCIONALIDADE E INSTRUMENTO 8

2.1 Estatísticas descritivas dos óbitos segundo a causa básica 11

3 METODOLOGIA 20

3.1 Estimação e interpretação dos modelos 22

4 HOMICÍDIOS OCULTOS NO BRASIL 34

4.1 Número projetado de homicídios 38

5 CONCLUSÃO 43

REFERÊNCIAS 44

SINOPSE

O Sistema de Informação sobre Mortalidade do Ministério da Saúde (SIM/MS) é uma das principais fontes de dados para se aferir mortes violentas e, em particular, homicídios no Brasil. No entanto, entre 1996 e 2021, ocorreram 294.752 óbitos violentos em que a causa não foi definida (ou 8,7% do total de mortes violentas do país). O objetivo deste estudo foi estimar a parcela desses óbitos que foram ocasionados por homicídios. A metodologia adotada no trabalho baseou-se em um modelo de aprendizado supervisionado (*machine learning*), em que o padrão probabilístico de características pessoais e situacionais para cada tipo de evento – se homicídio ou se suicídio/acidente – foi aprendido. Nossas estimativas indicaram haver no período 128.567 homicídios ocultos, ou homicídios não registrados oficialmente como tal. Segundo nossos achados, em média, a cada ano, 4.492 homicídios deixaram de ser registrados. Esse problema parece ser mais grave em quatro Unidades Federativas, a saber: São Paulo, Rio de Janeiro, Bahia e Minas Gerais.

Palavras-chave: aprendizado de máquina; subnotificação de homicídio; Sistema de Informação sobre Mortalidade (SIM); Brasil.

ABSTRACT

The Ministry of Health's Mortality Information System (SIM) is one of the main sources of information for measuring violent deaths and, in particular, homicides in Brazil. However, between 1996 and 2021, there were 294,752 violent deaths in which the cause was not defined (or 8,7% of the country's total violent deaths). The aim of this article was to estimate the proportion of these deaths that were caused by homicides. The methodology adopted in the work was based on a supervised learning model (machine learning), in which the probabilistic pattern of personal and situational characteristics for each type of event – whether homicide or suicide/accidents – was learned. Our estimates indicated that in the period there were 128,567 hidden homicides, or homicides not officially registered as such. According to our findings, an average of 4,492 homicides went unreported each year. This problem seems to be more serious in four Federal Units: São Paulo, Rio de Janeiro, Bahia and Minas Gerais.

Keywords: machine learning; homicide underreporting; Mortality Information System (SIM); Brazil.

1 INTRODUÇÃO

A taxa de homicídios por 100 mil habitantes tem sido utilizada como o principal indicador da criminalidade, não apenas em função da gravidade do dano, mas também pela maior disponibilidade da informação, e ainda porque esse tipo de incidente possui menor taxa de subnotificação se comparado a outros crimes, como assinalado nos trabalhos de Pinotti (2020) e Tabarrok, Heaton e Helland (2010).

Devido à heterogeneidade metodológica entre as agências responsáveis por informações criminais, pesquisadores e interessados na medição da criminalidade violenta encontram no Sistema de Informação sobre Mortalidade do Ministério da Saúde (SIM/MS) uma fonte de dados confiável e com metodologia transparente que justifica a sua utilização.

Os óbitos registrados no SIM podem ser divididos em mortes naturais e mortes por causas externas (ou violentas). Por sua vez, as mortes violentas são classificadas, segundo a 10ª revisão da Classificação Internacional de Doenças (CID-10), de acordo com as suas causas básicas,[1] que podem ser agregadas nos óbitos decorrentes de acidentes, suicídios, agressões, intervenções de agentes do Estado, complicações de assistência médica e mortes violentas cuja intenção foi indeterminada. Estudiosos em crime tipificam como homicídios a agregação de agressões e intervenções de agentes do Estado[2] (Cerqueira e Soares, 2016), conforme preconizado pelo Protocolo de Bogotá.[3]

1. Doença ou lesão que iniciou a cadeia de acontecimentos patológicos que conduziram diretamente à morte, ou as circunstâncias do acidente ou violência que produziram a lesão fatal (Brasil, CFM e CBCD, 2009). É codificada a partir da declaração do médico atestante, segundo regras estabelecidas pela Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde (CID-10), publicada pela Organização Mundial da Saúde (OMS).

2. Note-se que a definição de homicídios aqui adotada se presta a fins de análise epidemiológica e não coincide com o crime de "homicídio" descrito no Código Penal, ainda que as duas tipologias tratem de um fenômeno que muitas vezes coincidem, mas não necessariamente, como, por exemplo, no caso do chamado latrocínio, que, no Código Penal, é uma forma qualificada do crime de roubo, descrito em seu art. 157.

3. "O homicídio se define, para o presente propósito, como a morte de uma pessoa causada por uma agressão intencional de outra(s). Nesse sentido, excluem-se os homicídios não intencionais, os acidentais e as tentativas de homicídio (...). Além disso, são consideradas as mortes por agressão cometidas por agentes públicos no exercício do seu dever profissional, mesmo quando sejam legais, bem como as mortes acontecidas no exercício da legítima defesa por parte de qualquer pessoa. Em consequência, essa definição de homicídio não está limitada pela tipificação legal, que varia de país para país e inclui com frequência diversos tipos penais, mas por um conceito geral que não depende da legalidade ou ilegalidade dos fatos. Esta opção maximiza a comparabilidade internacional, é consistente com o objetivo de minimizar as mortes por agressão independentemente da sua legalidade e evita a demora que resulta da espera pela certeza de uma decisão judicial" (Open Society Foundations *et al*., 2015, p. 4).

No entanto, existe a possibilidade de parcela das mortes violentas que foram classificadas como mortes violentas de intenção indeterminada se tratar, na realidade, de homicídios ocultos. Com efeito, entre 1996 e 2021, 8,7% das mortes por causas externas foram classificadas como morte violenta por causa indeterminada (MVCI),[4] isto é, em média, o sistema de saúde não conseguiu identificar a intencionalidade de 11.336 óbitos ao ano. Em determinadas Unidades da Federação (UFs), o número de MVCIs ultrapassou o número de homicídios, impedindo a compreensão do real nível de violência. Por exemplo, em São Paulo, entre 2018 e 2021, o número de MVCIs foi, em média, 27,9% superior ao número de homicídios.

Sobre esse tema, analisando os dados de um conjunto de ex-repúblicas soviéticas, Värnik *et al*. (2010) sugerem a utilização das MVCIs como tentativa de ocultar o número de suicídios e homicídios. Na Rússia, de acordo com Andreev *et al*. (2015), entre 2000 e 2011, 33% das MVCIs foram, na realidade, homicídios incorretamente registrados, enquanto, na Finlândia, Ohberg e Lonnqvist (1998) sugerem que 10% das MVCIs são suicídios. Na Argentina, entre 1997 e 2018, segundo Santoro (2020), 28,5% das MVCIs foram identificadas como homicídios ocultos.

No caso brasileiro, analisando o município de Viçosa-MG, entre 2000 e 2009, Melo *et al*. (2014) identificaram que 36,8% dos 104 óbitos não registrados no SIM são homicídios não registrados, e 57,7% dos 391 óbitos por causa externa indeterminada são, na realidade, homicídios. No estado do Rio de Janeiro, Lopes *et al*. (2018) indicam fragilidades no fluxo de informação entre os responsáveis por registrar os óbitos e aqueles responsáveis em determinar intencionalidades das mortes. Por fim, utilizando a abordagem bayesiana e dados de 2001, Cavalini e Leon (2007) sugerem sub-registro de 5,9% de óbitos. Portanto, em diversos contextos, a literatura sugere subenumeração de óbitos por causas externas, isto é, o não registro da morte por causa externa[5] ou o registro dessa morte em categorias mal definidas, como MVCIs.

Cerqueira (2012; 2013) desenvolveu um método para atribuir a probabilidade a cada MVCI de ser, na realidade, um homicídio mal classificado. A partir daí, calculou a esperança matemática da ocorrência de homicídios ocultos por UF no Brasil em cada ano. O método basicamente identificava padrões diferenciados de mortes, segundo as características situacionais do incidente e características individuais das vítimas. Com base nos microdados dos óbitos, o autor aplicou modelos estatísticos dos tipos

4. Entre 1996 e 2021, foram registrados 294.736 MVCIs, ou mortes violentas em relação às quais o Estado não foi capaz de explicitar por que o cidadão morreu.

5. A identificação de óbitos não registrados no SIM está além do escopo deste trabalho. Para mais detalhes, ver Frias *et al*. (2017) e Diógenes *et al*. (2022).

multinomial e logit para reclassificar os óbitos com causa indeterminada como homicídios, suicídios ou acidentes.

Neste trabalho, seguimos a linha investigativa de Cerqueira (2013), e não apenas atualizamos o mapa dos homicídios ocultos até 2021, mas aprimoramos o método estatístico, incluindo técnicas de *machine learning*.

Após esta introdução, a seção 2 apresenta inicialmente a tipologia das agregações dos óbitos segundo a causa básica e o instrumento que gerou o primeiro processo mórbido. Nessa seção, apontamos também as estatísticas descritivas associadas aos óbitos violentos. Na seção 3, desenvolvemos a metodologia adotada no trabalho, bem como descrevemos os principais resultados, incluindo o número de homicídios ocultos e a taxa projetada de homicídios para cada UF. Em seguida, expomos as conclusões e discussões sobre políticas públicas.

2 DEFININDO AS TIPOLOGIAS DE ÓBITOS POR INTENCIONALIDADE E INSTRUMENTO

Por força de lei, nenhum sepultamento deveria ocorrer na ausência de certidão de óbito (Brasil, 1973). No caso de óbitos não naturais (mortes por causa externa, ou mortes violentas), as regulamentações Brasil (2009) e CFM (2005) tornam obrigatório o fornecimento, pelo serviço médico legal, de declaração de óbito (DO), informando a causa básica da morte.

A DO utilizada nos registros de morte por causa externa do SIM deve ser emitida por médico-legista, após laudo pericial cadavérico.[6] A partir desse exame pericial e de informações prestadas por familiares, indivíduos que socorreram a vítima ou pela polícia, o médico-legista tenta estabelecer a causa básica da morte, isto é , se a morte por causa externa foi ocasionada por: i) acidentes (por exemplo, acidente de carro); ii) lesões autoprovocadas intencionalmente (suicídio); iii) agressões (por exemplo, agressão por meio de arma de fogo); iv) intervenções legais e operações de guerra; v) MVCIs; ou vi) complicações de assistência médica, sequelas médicas ou acidente natural. Com base nas informações apuradas pelo médico-legista, os codificadores das secretarias municipais e estaduais de saúde irão preencher o código da CID-10 subjacente ao óbito. Na impossibilidade de distinguir entre homicídio, acidente e suicídio, a causa básica é classificada como MVCI.

6. Para uma descrição do fluxo de emissão da declaração de óbito, ver Brasil (2022).

A investigação do número de homicídios ocultos utiliza informações registradas nas DOs por causas externas, acessadas por meio dos microdados do SIM/MS, referentes ao período 1996-2021. Considerando-se óbitos pertencentes ao capítulo XX da CID-10 (Causas externas de morbidade e de mortalidade), os registros foram agregados em quatro grupos de óbito, apresentados no quadro 1, conforme a intencionalidade da causa básica do óbito e o instrumento que iniciou o processo mórbido: i) morte de intenção determinada resultante de acidente não natural ou suicídio são classificadas no grupo *acidentes/suicídios*; ii) as mortes de intenção determinada resultante de agressões ou intervenções legais são classificadas no grupo *homicídios*; iii) MVCIs são classificadas no grupo de *indeterminadas*; e iv) mortes de intenção determinada resultante de sequela médica, privação alimentar e acidentes naturais, tais como mordida de crocodilo, erupção vulcânica e vítima de raio, são excluídas da análise anterior, dada a impossibilidade desse tipo de óbito ser representado por MVCI.

O código CID-10, registrado na causa básica do óbito, informa o instrumento responsável por desencadear o primeiro processo mórbido. Com base na classificação original, criamos uma variável *instrumento*, em que agregamos, em dez categorias, as várias possibilidades existentes: envenenamento; enforcamento; afogamento; perfuração por arma de fogo (PAF); instrumento impactante; fogo; instrumento perfurante; instrumento contundente; instrumento desconhecido; e veículo.

Os óbitos causados por impacto resultam de uma variedade de eventos que incluem quedas, objetos em queda, esmagamento em contato com ferramentas e utensílios, explosão de caldeira e de outros materiais. A categoria *instrumento perfurante* inclui mortes ocasionadas por objetos perfurantes ou cortantes. O *instrumento contundente* inclui variedade de ações, como golpe, pancada, pontapé e mordedura. O *enforcamento* também inclui casos de estrangulamento. Em *fogo* se incluem os óbitos ocasionados por inalação de fumaça por consequência de fogo e incêndio. Os *envenenamentos* decorrem da ingestão de grande variedade de substâncias, como álcool, drogas psicoativas, medicamentos e solventes.

O quadro 1 exibe os grupos de óbitos classificados conforme a intencionalidade, o instrumento responsável pela causa básica do óbito e o código CID-10 considerado. Então, por exemplo, óbito acidental de pedestre em colisão com automóvel (CID-10:V03) é acidente não natural de intenção determinada e, portanto, integra o grupo *acidente/suicídio*, causado pelo instrumento *veículo*.

QUADRO 1

Classificação dos óbitos quanto à intenção e ao instrumento

Intencionalidade do óbito	Instrumento	Descrição da categoria CID-10	Código CID-10
Determinada resultante de acidente não natural ou suicídio – acidente/suicídio	Veículo	Acidente de transporte	V01-V99
	PAF	Exposição a forças mecânicas inanimadas	W32-W34
	Perfurante	Exposição a forças mecânicas inanimadas	W25-W26
	Impacto	Quedas	W00-W19
		Exposição a forças mecânicas inanimadas	W20-W24; W27-W31; W35-W43; W49
	Contundente	Exposição a forças mecânicas animadas	W51; W50
	Afogamento	Afogamentos e submersões acidentais	W65-W74
	Enforcamento	Outras ameaças acidentais à respiração	W75-W76
	Fogo	Exposição à fumaça, ao fogo e às chamas	X00-X09
	Envenenamento	Envenenamento acidental por exposição a substâncias nocivas	X40-X49
	Desconhecido	Exposição acidental a outros fatores não especificados	X58-X59
	Envenenamento	Lesões autoprovocadas voluntariamente	X60-X69
	Enforcamento		X70
	Afogamento		X71
	PAF		X72-X74
	Impacto		X75; X80-X81;
	Veículo		X82
	Fogo		X76-X77
	Perfurante		X78
	Contundente		X79
	Desconhecido		X83-X84
Determinada resultante de agressões e intervenções legais – homicídio	Envenenamento	Agressões	X85-X90
	Enforcamento	Agressões	X91
	Afogamento	Agressões	X92
	PAF	Agressões	X93-X95
		Intervenção legal	Y350
	Impacto	Agressões	X96; Y01-Y02
		Intervenção legal	Y351
	Veículo	Agressões	Y03
	Fogo	Agressões	X97-X98
		Intervenção legal	Y352
	Perfurante	Agressão	X99
		Intervenção legal	Y354
	Contundente	Agressões	Y00; Y04-Y05
		Intervenção legal	Y353
	Desconhecido	Agressões	Y06-Y09
		Intervenção legal	Y356-Y357

(Continua)

(Continuação)

Intencionalidade do óbito	Instrumento	Descrição da categoria CID-10	Código CID-10
Indeterminada – MVCI	Envenenamento	Eventos (fatos) cuja intenção é indeterminada	Y10-Y19
	Enforcamento	Eventos (fatos) cuja intenção é indeterminada	Y20
	Afogamento	Eventos (fatos) cuja intenção é indeterminada	Y21
	PAF	Eventos (fatos) cuja intenção é indeterminada	Y22-Y24
	Impacto	Eventos (fatos) cuja intenção é indeterminada	Y25; Y30; Y31
	Veículo	Eventos (fatos) cuja intenção é indeterminada	Y32
	Fogo	Eventos (fatos) cuja intenção é indeterminada	Y26-Y27
	Perfurante	Eventos (fatos) cuja intenção é indeterminada	Y28
	Contundente	Eventos (fatos) cuja intenção é indeterminada	Y29
	Desconhecido	Eventos (fatos) cuja intenção é indeterminada	Y33-Y34
Sequela médica ou acidentes naturais – excluído		Todas as outras causas externas	W44-W46; W52-W65; W77-W99; X10-X39; X50-X57; Y40-Y89; Y90-Y98

Fonte: CID-10.
Elaboração dos autores.

2.1 Estatísticas descritivas dos óbitos segundo a causa básica

Ao serem excluídos óbitos por sequelas médicas, privações e acidentes naturais, as mortes por causas externas consideradas totalizaram 3.396.010, entre 1996 e 2021. A seguir, apresentaremos as estatísticas descritivas associadas aos três conjuntos de causa básica do óbito: "homicídios", "suicídios e acidentes" e "MVCI". Nessa descrição, perceberemos que existem padrões distintos de mortalidade, a depender da causa do óbito.

A análise posterior, baseada nas técnicas de *machine learning*, objetiva, em última instância, aprender esses padrões diferenciados de letalidade, de modo a se estimar a probabilidade de as MVCIs terem sido, na realidade, homicídios.

A tabela 1 apresenta, nas colunas, a distribuição do número de óbitos em que se utilizou determinado instrumento para cada causa básica do óbito. Na tabela, cada linha mostra ainda como os óbitos para cada instrumento se distribuíram percentualmente entre as diferentes causas básicas do óbito.

Pode-se observar que 52,3% dos óbitos foram resultados de acidentes ou suicídios, sendo os quatro instrumentos mais prevalentes nessa causa básica, respectivamente, veículo, impacto, enforcamento e afogamento. Os homicídios respondem por 39,0% do total de óbitos, e os instrumentos mais utilizados foram, respectivamente, PAF, perfurante, desconhecido e contundente. As MVCIs responderam por 8,7% do total de mortes, sendo que instrumento desconhecido, PAF, instrumento contundente e enforcamento foram os meios mais utilizados nessa causa de mortalidade.

Ainda, é interessante perceber como parcela significativa dos óbitos relacionados a determinado instrumento se concentra em causas específicas de mortalidade violenta. Assim, 90,1% das mortes por afogamento estavam relacionadas a óbitos ocasionados por acidentes e suicídios. Da mesma forma, 93,2% dos óbitos por instrumentos perfurantes e 92,5% dos óbitos ocorridos mediante o uso de armas de fogo (PAF) ocorreram em casos de homicídio.[7]

TABELA 1
Instrumentos e intencionalidade do óbito (1996-2021)

Instrumento	Intenção							
	Homicídio		Acidente/suicídio		MVCI		**Total**	
	Quantidade	(%)	Quantidade	(%)	Quantidade	(%)	**Quantidade**	**(%)**
Afogamento	1.526	0,9	152.320	90,1	15.268	9,0	**169.114**	**100,0**
Contundente	87.766	71,4	2.966	2,4	32.131	26,2	**122.863**	**100,0**
Desconhecido	88.481	25,7	97.082	28,2	158.671	46,1	**344.234**	**100,0**
Enforcamento	17.295	9,0	158.163	82,6	15.912	8,3	**191.370**	**100,0**
Envenenamento	909	1,5	46.644	75,6	14.134	22,9	**61.687**	**100,0**
Fogo	5.497	12,8	30.871	72,1	6.430	15,0	**42.798**	**100,0**
Impacto	861	0,3	284.388	97,4	6.581	2,3	**291.830**	**100,0**
PAF	930.255	92,5	38.328	3,8	36.927	3,7	**1.005.510**	**100,0**
Perfurante	191.527	93,2	5.545	2,7	8.351	4,1	**205.423**	**100,0**
Veículo	1.860	0,2	958.974	99,8	347	0,0	**961.181**	**100,0**
Total	**1.325.977**	**39,0**	**1.775.281**	**52,3**	**294.752**	**8,7**	**3.396.010**	**100,0**

Fonte: MS/SVS/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Obs.: Os percentuais dizem respeito à distribuição para cada tipo de instrumento entre as diferentes causas básicas, isto é, homicídios, acidentes/suicídios e MVCI.

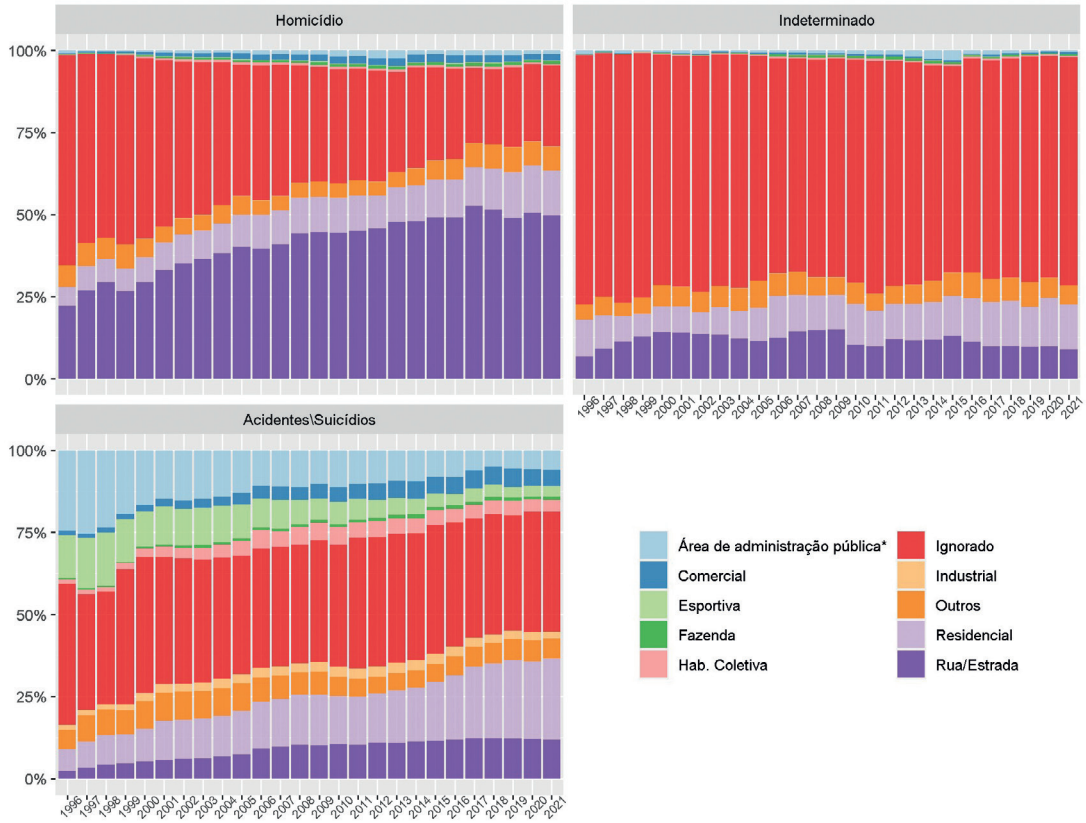
A seguir, apresentaremos a série temporal da distribuição de cada característica associada a determinada causa básica do óbito.

O local de ocorrência do incidente que gerou o óbito foi identificado com base no terceiro dígito[8] da CID-10. O gráfico 1 apresenta a distribuição do local do incidente. *Local desconhecido* é a informação de maior frequência nas MVCIs, representando uma média anual de 68,9% dos registros. No caso de homicídio, apesar da gradual redução temporal nos registros com local desconhecido, essa é a categoria com a segunda maior participação média nesses óbitos, com índice de 38,7% dos casos. Como se pode

7. A tabela A.1 do apêndice A apresenta a distribuição dos instrumentos, desagregando os grupos de óbitos.
8. No caso de intervenções legais (CID10:Y35) e operações de guerra (CID10:Y36), o terceiro dígito indica o instrumento, e não o local do incidente. Nesses casos, considerou-se o local do incidente como rua/estrada.

observar, ao longo do tempo, a categoria *local desconhecido* foi cedendo espaço aos registros de homicídios perpetrados nas ruas/estradas, que aumentou de uma participação de 22,1%, no primeiro período, para 49,7% em 2021. A parcela de registros de homicídios em residências também variou de 5,8% para 13,7%, no mesmo intervalo de tempo. Similarmente aos dois grupos de causa de óbito apontadas acima, no conjunto dos acidentes e suicídios, o local desconhecido representa parcela significativa dessas mortes, com média anual de 37,7%. No entanto, ao contrário do que ocorre nos casos de homicídio ou MVCI, locais como áreas de comércio, administração pública, prática esportiva e habitação coletiva representam parcela significativa dos óbitos envolvendo suicídios e acidentes, conforme o gráfico 1.

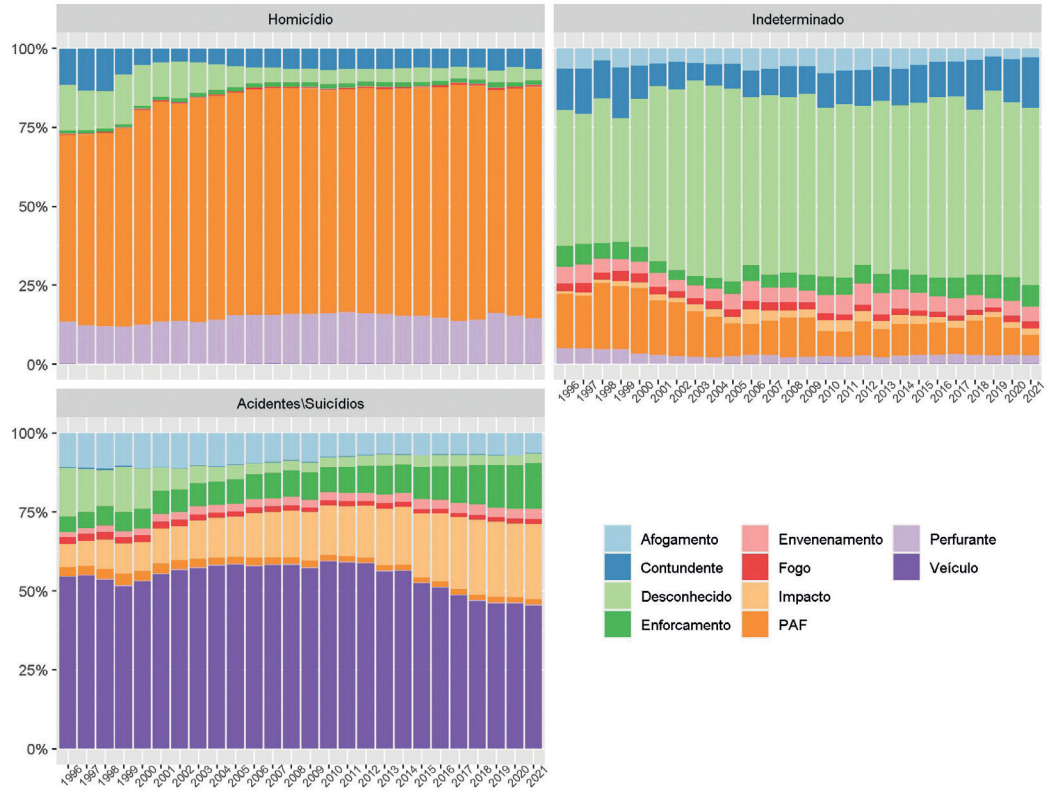
GRÁFICO 1
Local do incidente – Brasil (1996-2021)
(Em %)



Fonte: MS/SVS/Dados/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/. Elaboração dos autores.
Nota: [1] Áreas sob administração pública se incluem: hospitais, escolas e edifícios (inclusive áreas adjacentes) utilizados pelo público em geral ou por um grupo particular de pessoas.
Obs.: Ilustração cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

O gráfico 2 apresenta a evolução temporal da distribuição dos instrumentos associados a cada causa básica do óbito. Em todos os períodos, *instrumento desconhecido* aparece como aquele de maior frequência no caso das MVCIs, correspondendo à média anual[9] de 53,5% dos casos, seguido por *PAF* (12,4%) e *instrumentos contundentes* (10,8%). Nas mortes ocasionadas por homicídios, PAF é o instrumento de maior recorrência, com tendência de representatividade crescente e média anual de 69,7%. Nessa causa de óbito, os instrumentos perfurantes aparecem em segundo lugar, com média anual de 14,3%. No grupo dos acidentes e suicídios, o instrumento *veículo* é o principal causador dos óbitos, com média anual de 54,1%, seguido do instrumento *impacto*, com média anual de 15,5%. Nesse grupo, a participação de PAF é marginal, com média anual de 2,2%.

GRÁFICO 2
Instrumento da causa básica do óbito – Brasil (1996-2021)
(Em %)

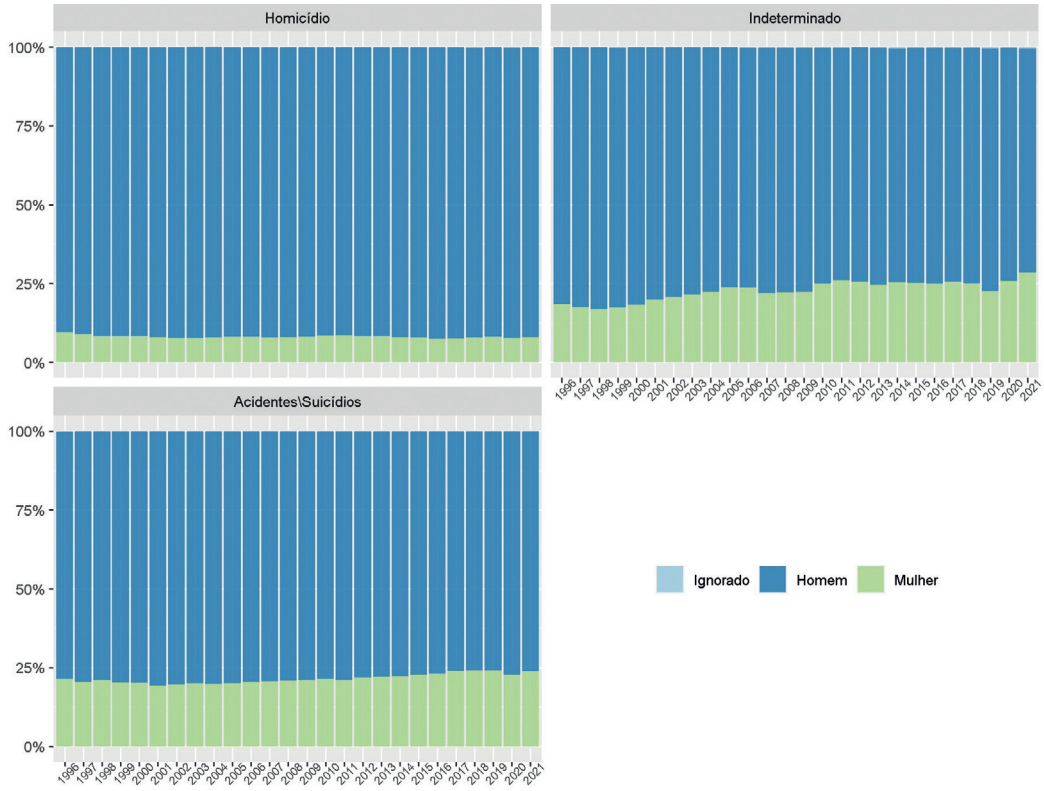


Fonte: MS/SVS/Dados/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Obs.: Ilustração cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

9. Note-se que esses números podem diferir dos explicitados na tabela 1, pelo fato de estarmos tratando aqui com médias anuais, e não com o percentual de casos no conjunto da amostra.

No que diz respeito ao sexo da vítima, os homens são as vítimas mais frequentes dos óbitos analisados, conforme apresentado no gráfico 3. Ao longo dos 24 anos analisados, homens foram vítimas de 91,8% dos homicídios, 77,0% das mortes por causa externa de intenção indeterminada e 78,4% dos acidentes/suicídios. No caso de sexo da vítima, a incerteza sobre a informação é residual, e, em nenhuma das intencionalidades analisadas, essa característica é ignorada em mais de 1% dos registros.

GRÁFICO 3
Sexo da vítima – Brasil (1996-2021)
(Em %)

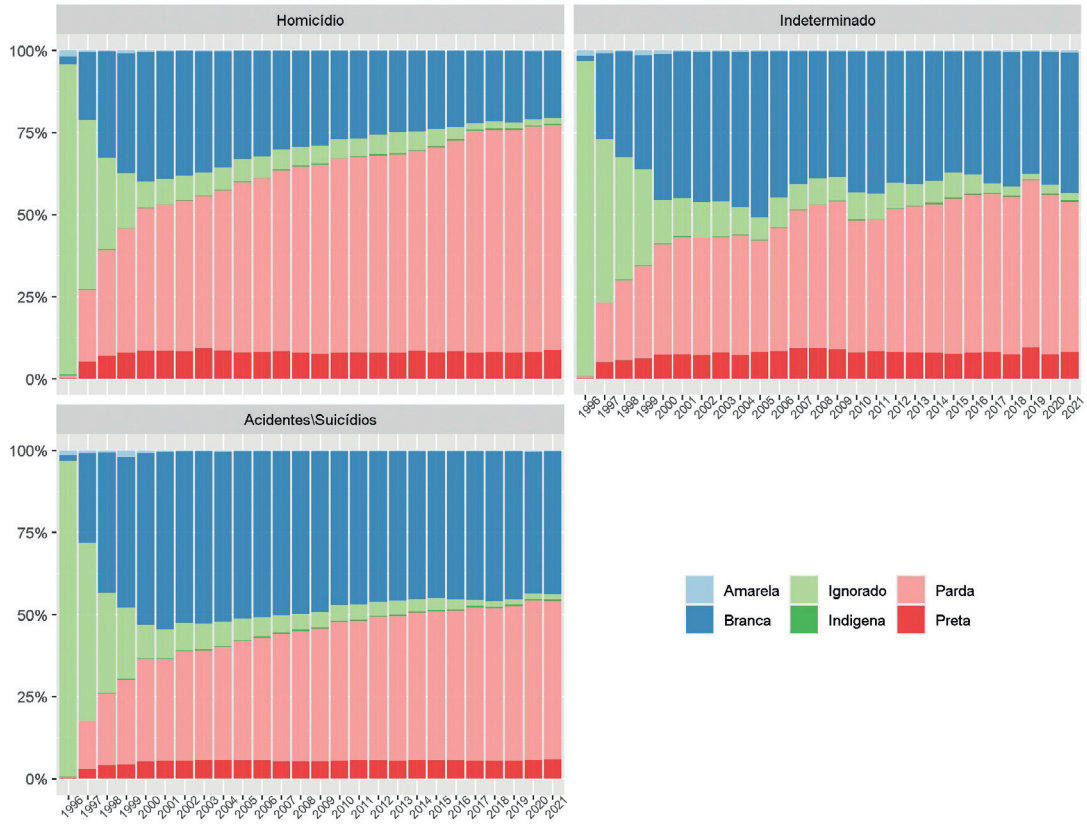


Fonte: MS/SVS/Dados/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/. Elaboração dos autores.
Obs.: Ilustração cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

No que diz respeito à raça/cor da vítima, o percentual de informação desconhecida apresenta contínua diminuição em todas as modalidades de intencionalidade dos óbitos, conforme o gráfico 4. No caso de homicídio, são mais frequentes as vítimas pardas, com participação crescente no total de vítimas, e média anual de 52,3%, seguidas de brancos, com 27,5%. A proporção de pretos apresenta estabilidade, com média anual de 7,7%. Em se tratando das MVCIs, os brancos são as principais vítimas, com média

anual de 39,0% dos óbitos, diferença marginalmente superior à média anual dos pardos, de 38,4%. No caso de pretos, ocorre estabilidade ao longo dos anos, em patamar similar aos casos de homicídio, e média anual de 7,4%. No grupo de óbitos composto por acidentes e suicídios, os brancos são as principais vítimas, com média anual de 45,0% dos óbitos, seguidos por vítimas pardas, com média de 37,1%, e vítimas de raça/cor preta, com média anual de 5,0%. Portanto, não parece ocorrer significativa diferença na distribuição de raça/cor nos óbitos analisados.

GRÁFICO 4
Raça/cor da vítima – Brasil (1996-2021)
(Em %)

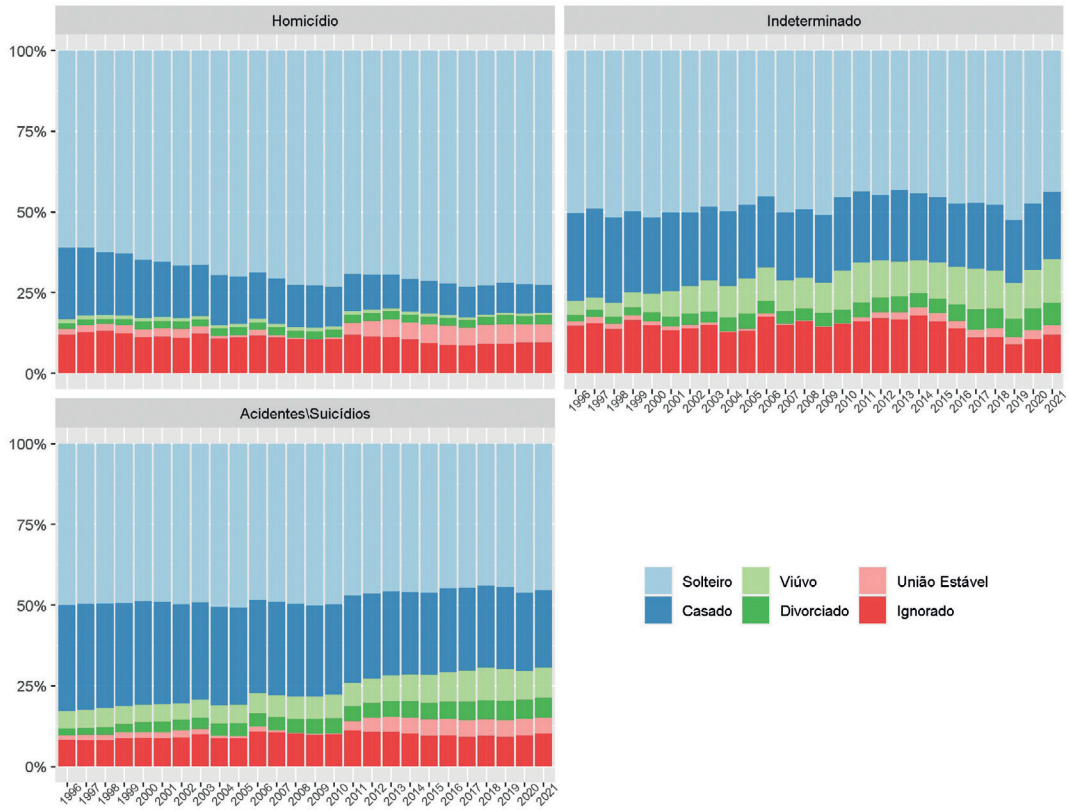


Fonte: MS/SVS/Dados/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Obs.: Ilustração cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

No que diz respeito ao estado civil, os dados no gráfico 5 indicam que os solteiros são as principais vítimas de homicídios, com média anual de 69,0%, seguidos pelos casados, com média anual de 13,6%, e vítimas com estado civil desconhecido, com média de 10,7%. Em relação ao conjunto dos acidentes e suicídios e às MVCIs, os solteiros são também as principais vítimas, com médias anuais de 47,8% e 48,0%,

respectivamente, seguidos pelo grupo de casados, com médias anuais de 28,4% e 22,3%, respectivamente. Nesses dois tipos de óbitos, os solteiros e casados cederam espaço aos viúvos e divorciados, que, aproximadamente, dobraram de participação. Ainda assim, o estado civil desconhecido compreende parcela relevante desses óbitos, representando, em média, 9,5% dos acidentes/suicídios e 14,3% das mortes por causa externa indeterminada.

GRÁFICO 5
Estado civil da vítima – Brasil (1996-2021)
(Em %)

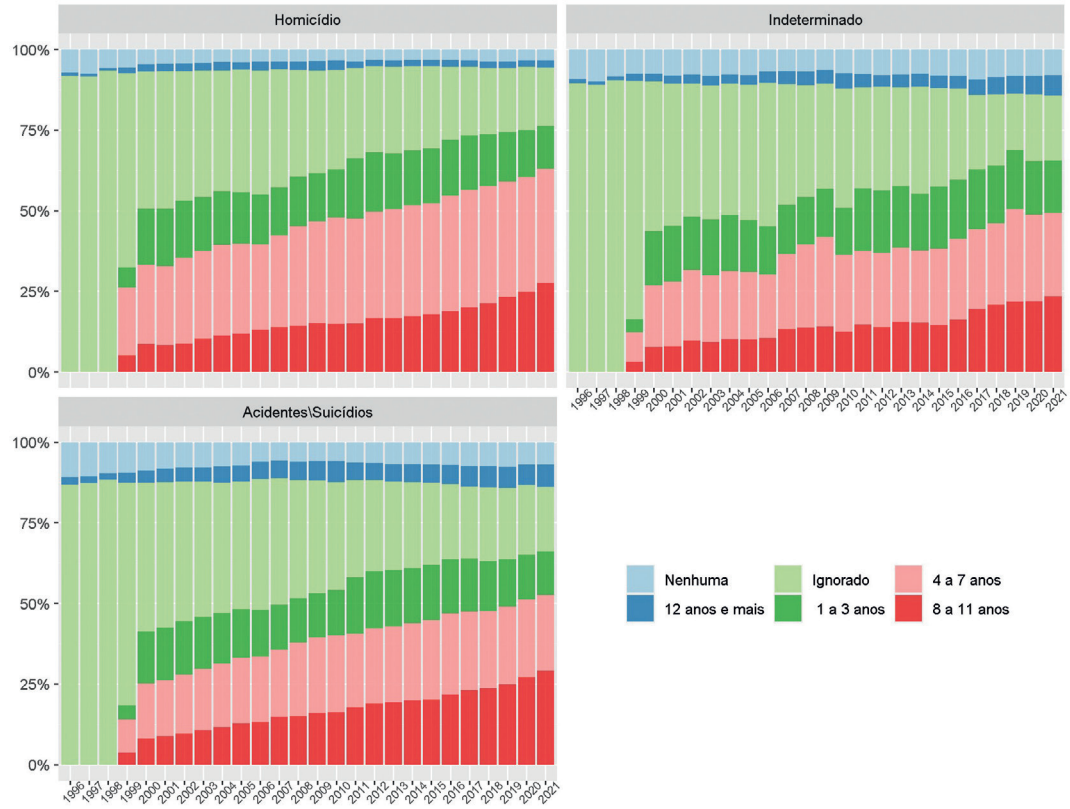


Fonte: MS/SVS/Dados/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Obs.: Ilustração cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

Em relação à escolaridade da vítima, a distribuição do nível educacional é aproximadamente homogênea entre os três tipos de óbitos analisados, conforme o gráfico 6. A escolaridade desconhecida apresenta gradativa diminuição, tal que, nos três tipos de óbitos, entre o início e o final do período analisado, reduziu-se sua participação em 70 pontos percentuais (p.p.), encerrando 2021 com aproximadamente 20% das vítimas em cada tipo de óbito – ainda assim, representando parcela relevante dos vitimizados.

Nos três tipos de óbitos, a redução de vítimas com escolaridade desconhecida foi causada pela expansão de três níveis de escolaridade, coletivamente abrangendo vítimas com escolaridade entre 1 e 11 anos. Nesse grupo, vítimas com escolaridade entre 4 e 7 anos são os óbitos mais frequentes, respondendo em média por 27,4% dos homicídios, 19,1% dos acidentes/suicídios, e 20,0% das mortes por causa externa de intenção indeterminada.

GRÁFICO 6
Escolaridade da vítima – Brasil (1996-2021)
(Em %)



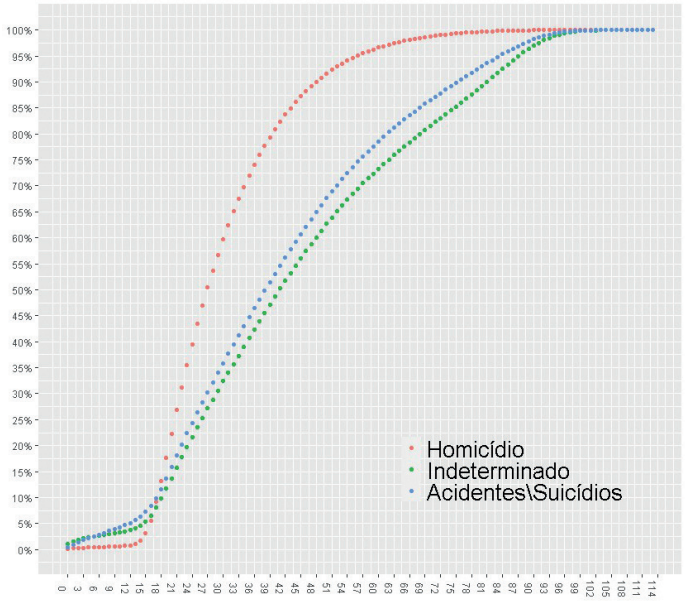
Fonte: MS/SVS/Dados/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/. Elaboração dos autores.
Obs.: Ilustração cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

Por fim, o gráfico 7 apresenta a função de distribuição acumulada empírica de idade da vítima nos três tipos de óbitos analisados. No caso de homicídios, vítimas com menos de 14 anos são pouco representativas; entretanto, após essa idade, o fardo da violência sobre os jovens é evidente, tal que 50% das vítimas de homicídios têm até 27 anos. Por outro lado, no caso de acidente/suicídio e MVCI, vítimas de até 14 anos representam aproximadamente 5% dos óbitos, e a vitimização é menos concentrada em jovens. Nesses dois grupos, aproximadamente 50% dos óbitos acontecem até os

40 anos. Portanto, parece existir distinção na distribuição de idade entre homicídios e demais óbitos, ocorrendo desproporcional acúmulo de jovens vítimas de homicídio.[10]

Em resumo, apesar de algumas similaridades, as distribuições de características indicam a existência de padrões distintos associados aos três tipos de óbitos. No que se refere a distribuição de raça/cor, estado civil, sexo da vítima e escolaridade, percebemos certa similaridade entre as categorias de morte violenta. Entretanto, em outras características, são registradas diferenças relevantes, especialmente no caso de instrumento gerador da causa básica de óbitos; nesse caso, cada tipo de óbito singulariza um instrumento majoritário. Ainda, como assinalado anteriormente, no caso da idade da vítima, os homicídios estão concentrados entre os mais jovens, enquanto, nos outros óbitos, ocorre maior dispersão entre as idades das vítimas. Por fim, a distribuição de local do incidente em acidente/suicídio apresenta diversificação inexistente nos outros óbitos, acontecendo as mortes em locais com reduzida ocorrência de homicídios e MVCIs, tais como áreas comerciais, de administração pública e industriais.

GRÁFICO 7
Idade da vítima – Brasil (1996-2021)
(Em %)



Fonte: MS/SVS/Dados/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Obs.: Ilustração cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

10. O gráfico A.1 do apêndice A apresenta o histograma da idade da vítima e o instrumento, por sexo da vítima e intencionalidade do óbito.

3 METODOLOGIA

Conforme descrito no quadro 1, a investigação considera três tipos de morte por causa externa: o grupo homicídios totais (H*) – que inclui as agressões e intervenções legais –, o grupo de acidentes e suicídios totais (AS*) e o grupo MVCI. Note-se que AS* e H* são apenas parcialmente observados, pois a porção não observada está classificada erroneamente como MVCI. As equações (1), (2) e (3) sintetizam o problema:

$$H^* = H + HO \tag{1}$$

$$AS^* = AS + ASO \tag{2}$$

$$\text{MVCI} = HO + ASO \tag{3}$$

Nas equações, *HO* e *ASO* são, respectivamente, homicídios ocultos e acidentes e suicídios ocultos, incorretamente registrados como MVCI. Em nossa notação, *H* e *AS* representam, respectivamente, os homicídios e os acidentes e suicídios registrados no SIM, cuja intenção foi determinada pelo sistema de saúde. O objetivo do trabalho é estimar o número de *HO* registrados no SIM como MVCI.

O não registro do homicídio no SIM (no caso em que o corpo da vítima não aparece), ou o registro de incidentes dessa natureza em classificações de mortes naturais, implica considerar que o número de homicídios ocultos encontrados aqui é um limite inferior para o número de homicídios que não foi registrado corretamente como tal. Por sua vez, desde que a causa da morte por causa externa tenha sido determinada, supõe-se que a classificação está correta.

Neste estudo, utilizamos métodos de classificação supervisionada para obter a predição dos HO. Trata-se de métodos heurísticos baseados em grande volume de dados e capacidade de processamento compatível, em que basicamente os algoritmos aprendem os padrões de ocorrência de determinado fenômeno e usam esse aprendizado para classificar o dado que possui alguma característica desconhecida, no caso, a causa identificada do óbito, se HO ou ASO, para as MVCIs.

Na análise em questão, o método revela padrões distintos de mortalidade, no que se refere às características das vítimas e elementos situacionais do incidente (instrumento usado na violência, local do incidente, ano, mês, dia da semana do óbito e UF de ocorrência) e probabilidades associadas a esses padrões. Para tornar mais clara a ideia, imaginemos uma MVCI cuja vítima era do sexo masculino, solteiro, negro, jovem, e foi morto na rua por uma arma de fogo. Tomando como base as estatísticas descritivas na seção 2.1, seríamos levados a crer que tal óbito seria resultante de um homicídio, e não de um suicídio ou acidente. Os métodos de classificação fazem a predição a partir de

uma análise muito mais refinada do que a mera observação de prevalência relativa dos casos, utilizando um processo intenso de reamostragem de grandes volumes de dados para identificar os padrões de mortalidade e associar probabilidades a esses padrões.

Neste trabalho, a identificação dos HO segue metodologia utilizada em diversos problemas de classificação (Nascimento *et al*., 2021) e acontece em três etapas: otimização de hiperparâmetros[11] (*model selection*), teste de generalização (*model assessment*) e ajuste do modelo final.

Na primeira etapa, a base de dados dos óbitos – composta pela variável dependente binária da causa do óbito (*H* ou *AS*) e por variáveis previsoras (todas as características da vítima e situacionais, conforme descritas anteriormente) – foi dividida de forma aleatória, estratificada em base de treinamento (70%) e base de teste (30%), de modo a se manter em cada partição a proporção de homicídios e acidentes/suicídios da base de dados original (*stratified random sampling*).[12] Então, nessa etapa da otimização de hiperparâmetros, utilizando-se 10-*fold cross-validation*, são ajustados sete algoritmos à base de treinamento, a saber:[13] *logistic regression*; *penalized logistic regression* (*ridge*, *lasso* e *elastic net*); *decision tree*; *bagging decision trees*; e *random forest*.[14] São diversas as possíveis combinações de hiperparâmetros, e nessa etapa busca-se identificar a combinação com melhor capacidade de previsão fora da amostra de cada algoritmo; por isso, a combinação de hiperparâmetros de maior média da área sob a curva ROC é selecionada.[15]

11. Os hiperparâmetros são atributos que controlam o ajuste do modelo de *machine learning*, sem o que não se pode garantir sua boa qualidade preditiva.

12. Por causa do desbalanceamento na base de treinamento, composta por 42,9% de homicídios, a base de treinamento foi reamostrada, utilizando-se *synthetic minority oversampling technique* (Smote). Em base desbalanceada, o viés de previsão da classe majoritária reduz a capacidade preditiva do algoritmo de classificação (Brodley e Friedl, 1999). A reamostragem Smote proposta por Chawla *et al*. (2002) equilibra a proporção de classes, ao expandir a base original através da interpolação aleatória de observações da classe minoritária com características semelhantes.

13. No quadro A.1 do apêndice A, apresentamos uma explicação sucinta de cada um desses algoritmos.

14. A capacidade computacional disponível aos pesquisadores determinou as etapas de pré-processamento, o número de classes consideradas na variável independente (duas) e os algoritmos utilizados. São apresentados detalhes dos algoritmos e métricas de desempenho preditivo nas tabelas B.1 e B.2 do apêndice B.

15. *Area under the ROC curve* (AUC-ROC). Do inglês, *receiver operating characteristic curve* (ROC curve), ou, simplesmente, curva ROC, é uma representação gráfica que ilustra o desempenho de um sistema classificador binário à medida que o seu limiar de discriminação varia, ou mostra a relação entre a parcela de casos classificados como verdadeiros positivos e falsos positivos.

A etapa de *model assessment* investiga possibilidade de *overfitting*[16] e capacidade de generalização da combinação de hiperparâmetros selecionados na etapa anterior, ao se realizar classificação da base de teste. O algoritmo e a respectiva combinação de hiperparâmetros de maior área sob a curva ROC na base de teste serão utilizados na identificação de homicídios ocultos.

Na etapa de identificação de homicídios ocultos, o algoritmo selecionado na etapa *model assessment* é ajustado em todo o conjunto de dados (treinamento e teste), estabelecendo-se o modelo preditivo utilizado (Murphy, 2020).

As variáveis preditivas são aquelas com informação ao longo de todo o período analisado. Assim, os previsores são majoritariamente variáveis categoriais, isto é, sexo, raça/cor, estado civil, escolaridade da vítima, local do incidente, instrumento da causa básica do óbito, ano, mês, dia da semana do óbito e UF de ocorrência. As variáveis categóricas são transformadas em variáveis binárias dos componentes da variável (*indicator variable*). A idade da vítima é a única variável contínua. Os *missing value*s em idade da vítima são imputados através do método *k-nearest neighbors* (k = 5), a partir das informações das mortes conhecidas e utilizando-se as demais variáveis com previsores da imputação.

Por fim, a interpretação dos modelos utiliza o *SHAP value* proposto por Lundberg e Lee (2017), ao verificar a contribuição das variáveis previsoras na probabilidade estimada, e o *partial dependence profile* de Biecek (2018), na compreensão da relação entre a probabilidade estimada de o óbito ser homicídio oculto e as características do óbito. A importância dos previsores no desempenho preditivo é analisada utilizando-se a *permutation feature importance* (PFI) de Fisher, Rudin e Dominici (2019).

3.1 Estimação e interpretação dos modelos

A tabela 2 apresenta a média das métricas de desempenho preditivo dos sete algoritmos, ao se aplicar 10-*fold cross-validation* na base de treinamento. Nessa etapa, os algoritmos são otimizados; portanto, em cada algoritmo, apresentamos a combinação de hiperparâmetros de maior área sob a curva ROC. No processo de *cross-validation* da base de treinamento, os algoritmos alcançaram excelente desempenho preditivo, e o *random forest* foi aquele de melhor desempenho, superando os demais algoritmos nas dez métricas consideradas.

16. *Overfitting*, ou ajuste excessivo, é um comportamento indesejável de aprendizado de máquina que ocorre quando o modelo fornece previsões precisas para dados de treinamento, mas não para novos dados.

TABELA 2
Métricas de *performance* no *cross-validation*: *model selection*

Modelos	Accuracy	Bal Accuracy	Precision	Sensitivity	Specificity	ROC AUC	F meas	J index	MCC
Random free	0,9623	0,9625	0,9484	0,9642	0,9608	0,9915	0,9562	0,9250	0,9231
Bagging tree	0,9598	0,9608	0,9395	0,9682	0,9534	0,9907	0,9536	0,9217	0,9184
DTree	0,9585	0,9597	0,9373	0,9677	0,9517	0,9895	0,9522	0,9193	0,9159
Logit	0,9573	0,9581	0,9377	0,9641	0,9521	0,9898	0,9507	0,9163	0,9133
Lasso logit	0,9572	0,9581	0,9377	0,9640	0,9522	0,9898	0,9507	0,9162	0,9132
Ridge logit	0,9539	0,9545	0,9356	0,9583	0,9507	0,9873	0,9468	0,9090	0,9064
Elastic net logit	0,9572	0,9581	0,9377	0,9640	0,9522	0,9898	0,9507	0,9162	0,9132

Elaboração dos autores.

Após a otimização de hiperparâmetros na base de treinamento, em cada algoritmo, a combinação de hiperparâmetros de maior AUC-ROC é utilizada na previsão da base de teste. Essa segunda rodada de previsões verifica a ocorrência de *overfitting* e investiga a capacidade de generalização dos algoritmos, isto é, a capacidade de acertadamente elaborar a classificação de observações inéditas; neste trabalho, a capacidade do algoritmo de identificar os homicídios ocultos.

De acordo com as evidências apresentadas na tabela 3, analogamente ao processo de *cross-validation* na base de treinamento, nas previsões de observações inéditas da base de teste, os algoritmos utilizados apresentam excelente desempenho preditivo. O bom desempenho na classificação de observações inéditas sugere adequada capacidade de generalização (Hastie, Tibshirani e Friedman, 2009; James *et al*., 2021; Kuhn e Johnson, 2013) e ausência de *overfitting*. Na etapa de teste, o *random forest*, por pequena margem, apresentou o melhor desempenho nas dez métricas consideradas e, portanto, foi selecionado como modelo a ser utilizado na identificação dos homicídios ocultos.[17]

Grosso modo, esse algoritmo *random forest* se inicia com a escolha de cem amostras aleatórias com reposição, do mesmo tamanho da amostra total de casos registrados como homicídios ou suicídios e acidentes, isto é, 3,1 milhões de observações. Para cada uma dessas amostras, é construída uma árvore de decisão, em que os nós e vértices são formados a partir da escolha aleatória das variáveis, até o nível em que o número de observações no nó situado no ramo mais inferior seja no mínimo igual a *min_n*, que é

17. Um teste formal poderia demonstrar a superioridade preditiva de algum modelo. Entretanto, por causa da similaridade de desempenho preditivo entre os modelos e falta de consenso sobre o teste a ser utilizado (Benavoli *et al*., 2017), optamos por utilizar o modelo de melhor desempenho na maioria das métricas.

um *hiperparâmetro* escolhido na primeira etapa de *model selection*. No nó final de cada caminho da árvore, computa-se o número de casos em que houve homicídio ou acidente e suicídio. A proporção de homicídios em relação ao número total representará a probabilidade frequentista de homicídios em cada um desses nós finais. Construídas essas árvores, as características associadas a cada MVCI se encaixarão em uma determinada regra de negócio, isto é, se seguirá um determinado caminho de cada árvore, sendo que no nó final constará a probabilidade de aquele óbito ser homicídio. A probabilidade predita de esse óbito ser homicídio será a média aritmética da probabilidade nas cem árvores. Caso essa probabilidade predita se situe acima de determinado valor *p** (no caso, *p**= 0,5, como veremos a seguir), aquela MVCI será classificada como homicídio.

TABELA 3
Métricas de *performance* no *test set*: *model assessment*

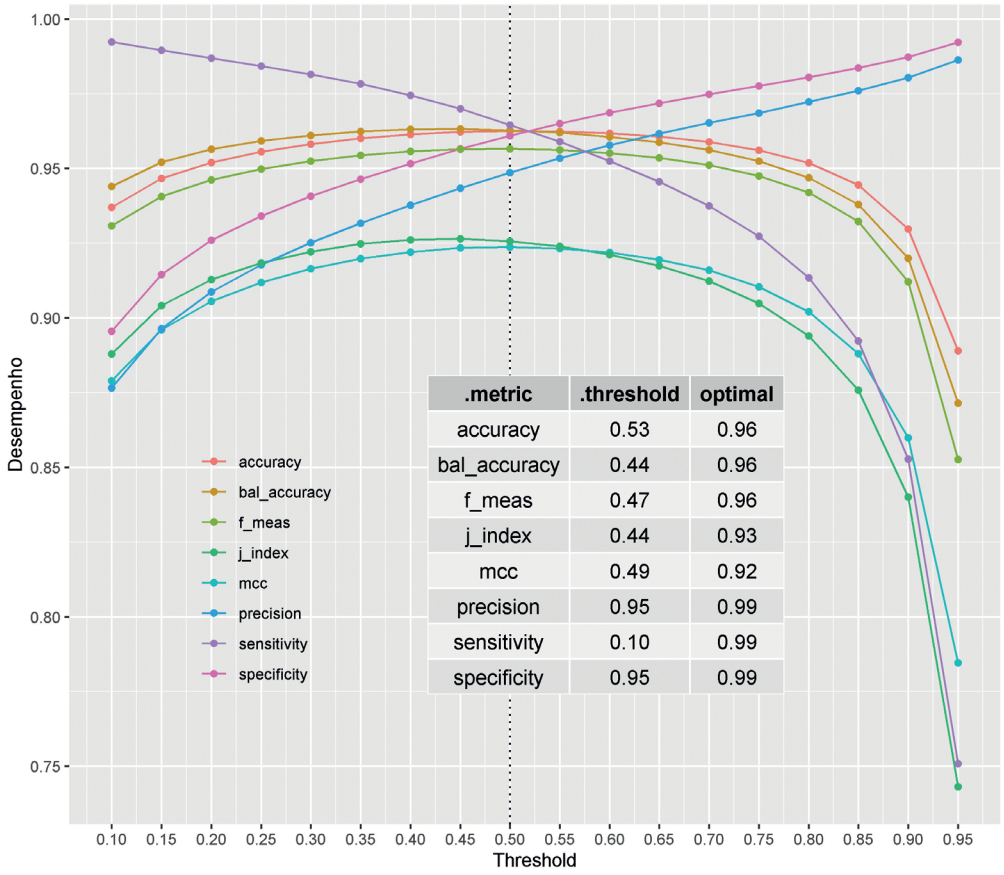
Modelos	Accuracy	Bal Accuracy	Precision	Sensitivity	Specificity	ROC AUC	F meas	J index	MCC	Pr_auc
Random free	0,9625	0,9628	0,9486	0,9646	0,9610	0,9917	0,9566	0,9256	0,9237	0,9871
Bagging tree	0,9464	0,9610	0,9405	0,9593	0,9510	0,9901	0,9469	0,9175	0,9130	0,9813
DTree	0,9590	0,9602	0,9380	0,9682	0,9522	0,9898	0,9529	0,9204	0,9170	0,9508
Logit	0,9575	0,9584	0,9378	0,9646	0,9522	0,9899	0,9510	0,9168	0,9138	0,9839
Lasso logit	0,9574	0,9583	0,9377	0,9644	0,9522	0,9899	0,9509	0,9166	0,9136	0,9839
Ridge logit	0,9541	0,9547	0,9355	0,9588	0,9507	0,9875	0,9470	0,9094	0,9068	0,9819
Elastic net logit	0,9574	0,9583	0,9377	0,9644	0,9522	0,9899	0,9509	0,9166	0,9136	0,9839

Fonte: MS/SVS/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.

A seguir, são examinados vários aspectos técnicos sobre o modelo *random forest* desenvolvido, bem como a importância das variáveis utilizadas para a predição, segundo esse modelo.

No gráfico 8, utilizando-se a base de teste, são apresentados os *thresholds* ótimos nas diversas métricas investigadas e as variações no desempenho preditivo, ao se alterar o *threshold* de classificação do modelo *random forest* utilizado na identificação dos homicídios ocultos. Observa-se que o *threshold* de 50% não é considerado o ponto ótimo em nenhuma das métricas. No entanto, entre as métricas que consideram as diferentes dimensões da matriz de confusão, o ponto ótimo está próximo ao limiar de 50% e apresenta ganho preditivo marginal. Por atribuirmos igual importância às diversas dimensões da matriz de confusão, e por causa dos ganhos marginais nos *thresholds* ótimos, o *threshold* de 50% foi utilizado na identificação dos homicídios ocultos. Entretanto, a depender do *threshold* ótimo, o número de homicídios ocultos poderia variar marginalmente.

GRÁFICO 8
Optimal thresholds



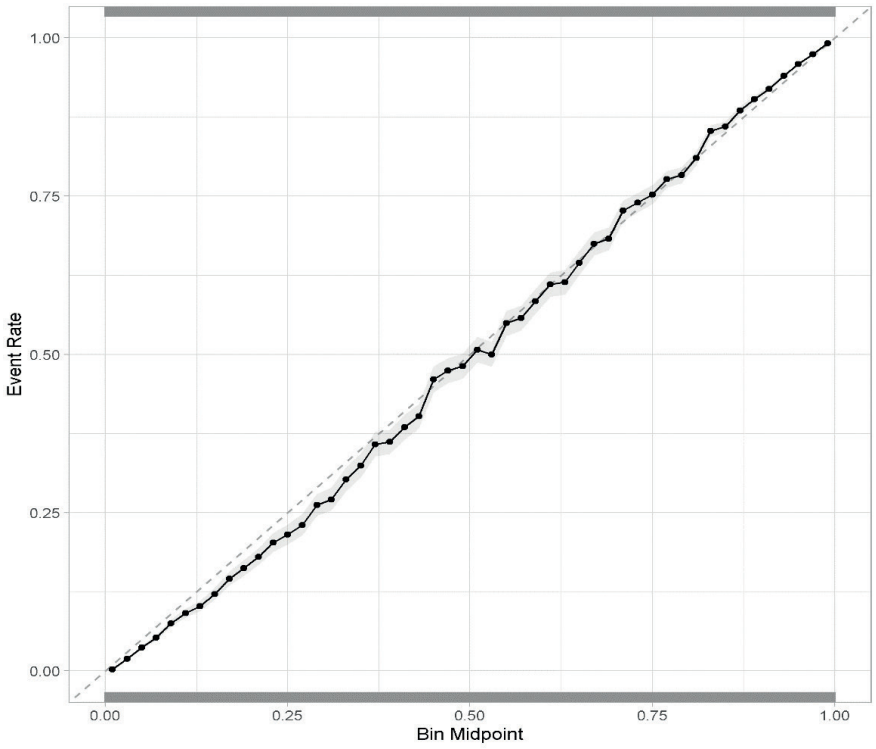
Fonte: MS/SVS/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Obs.: 1. Com base no modelo *random forest*.
2. Ilustração cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

Por fim, em virtude do desbalanceamento entre as intencionalidades, é importante verificar se o modelo utilizado está bem balanceado, isto é, se as probabilidades estimadas refletem a ocorrência observada entre as classes. Por exemplo, se coletarmos um grupo de observações cujas probabilidades estimadas são 60% de serem homicídios, então devemos esperar que cerca de 60% das observações no grupo sejam, de fato, homicídios. Um modelo mal calibrado pode apresentar adequado desempenho preditivo e, ainda assim, estar superestimando ou subestimando consistentemente as probabilidades reais. Por exemplo, se o modelo está subestimando as probabilidades reais, existe evidência de dificuldade em identificar os homicídios ocultos.

O gráfico 9 apresenta uma curva de calibração (*calibration plot*) no *test set*. Os valores das probabilidades estimadas de o óbito ser homicídio são agrupados em

cinquenta intervalos discretos ou *bins*. Em cada *bin*, calculamos a frequência de homicídios observados nos dados de teste. Por exemplo, em um *bin* com previsões de probabilidade entre 0,1 e 0,2, verificamos quantos homicídios ocorreram nesse intervalo. Nesse gráfico, o eixo x representa a média das probabilidades em cada *bin*, enquanto o eixo y mostra a frequência de homicídios observados. Esperamos que os pontos no gráfico sigam uma linha diagonal, sugerindo calibração perfeita. Quando os pontos estão acima dessa curva, o modelo está subestimando a probabilidade real, e, se estiverem abaixo da curva, o modelo estará superestimando a probabilidade real. A curva de calibração apresenta marginal superestimação até o *bin* 22; em seguida, a maioria dos pontos segue aproximadamente a linha de calibração ideal. Isso indica que, em média, as previsões estão bem calibradas e refletem a frequência de eventos observados.

GRÁFICO 9
Curva de calibração: *test set*

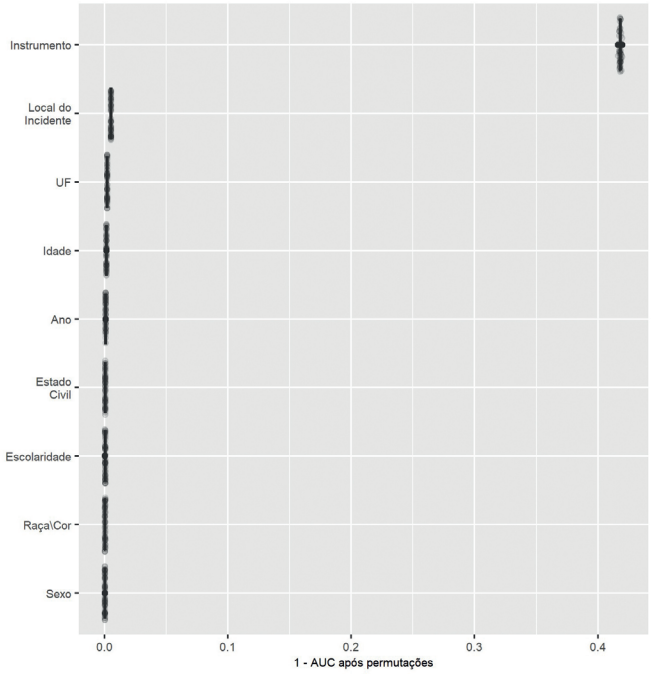


Fonte: MS/SVS/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Obs.: 1. Com base no modelo *random forest*.
2. Ilustração cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

Antes da análise dos homicídios ocultos, precisamos interpretar a estimação do modelo utilizado, isto é, entender a influência das variáveis preditivas nas probabilidades estimadas e no desempenho preditivo do modelo.

O gráfico 10 apresenta o PFI do *random forest* na base de teste. Proposto por Fisher, Rudin e Dominici (2019), o PFI, ao eliminar a relação entre variáveis preditivas e variável prevista, através de permutação no valor das variáveis preditivas, aufere a importância dos previsores no desempenho preditivo do modelo, atribuindo maior relevância aos previsores de elevado impacto no desempenho preditivo. Assim, seja *X* a matriz de previsores, o erro do modelo estimado $\widehat{f(X)}$ corresponde à função de perda $l^{orig} = (Y, \widehat{f(X)})$, em que *Y* representa o vetor de valores observados. Então, de forma iterativa, os valores de cada variável *j* são permutados e em seguida é calculado o valor da função de perda da variável permutada *j*, $l_j^{perm} = (Y, f(\widehat{X_j^{perm}}))$.[18] A importância da variável *j* será a diferença entre função de perda permutada e original, isto é, $VI_j = l_j^{perm} - l^{orig}$. Portanto, quanto maior a diferença no desempenho preditivo após a permutação, maior a importância da variável.

GRÁFICO 10
Permutatuion feature importance



Fonte: MS/SVS/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Obs.: 1. Por brevidade, somente as oito variáveis mais importantes são apresentadas.
2. Com base no modelo *random forest*.
3. Ilustração cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

18. Por causa da aleatoriedade introduzida ao se permutar a variável *j*, são realizadas cinquenta permutações em cada variável. A métrica considerada é 1-AUC.

O instrumento responsável pela causa básica do óbito foi a variável de maior importância e, portanto, aparece como importante previsor das duas classes investigadas, com redução aproximada de 0,418 na métrica AUC, seguido por local do incidente, com redução de 0,005; as demais variáveis são pouco influentes no desempenho preditivo. Como visto na seção anterior, o instrumento do óbito e o local do incidente são as variáveis com maior diferença, entre homicídio e acidente/suicídio, na distribuição das categorias constitutivas da variável, enquanto as demais variáveis apresentam similar distribuição de características. Portanto, a evidência de elevado impacto na capacidade preditiva causado pela variável *instrumento do óbito* reforça a importância do correto preenchimento da DO, já destacado em outros trabalhos (Soares Filho, Cortez-Escalante e França, 2016).

O *permutation based feature importance* identifica a importância das variáveis no desempenho preditivo do modelo, entretanto, não explica a influência das variáveis preditivas nas probabilidades estimadas. O método *shapley additive explanations* (*SHAP*) proposto por Lundberg e Lee (2017) e Aas, Jullum e Loland (2021) explica a probabilidade estimada em cada observação, ao atribuir a cada variável preditiva a contribuição percentual dessa variável na probabilidade estimada, indicando o sentido e a intensidade da associação entre o valor das variáveis previsoras e a probabilidade estimada. Inspirado no *Shapley value* de Shapley (1953), o método foi originalmente idealizado com objetivo de repartir recompensa de jogo cooperativo entre participantes de coalização. No contexto de *machine learning*, a probabilidade estimada de cada observação será repartida entre as *P* variáveis preditivas, tal que o *SHAP* da *j*-ésima variável é média ponderada da contribuição marginal da variável *j* na probabilidade estimada em todas as possíveis coalizões de variáveis *p!*. Formalmente, o *SHAP* da *j*-ésima variável será:

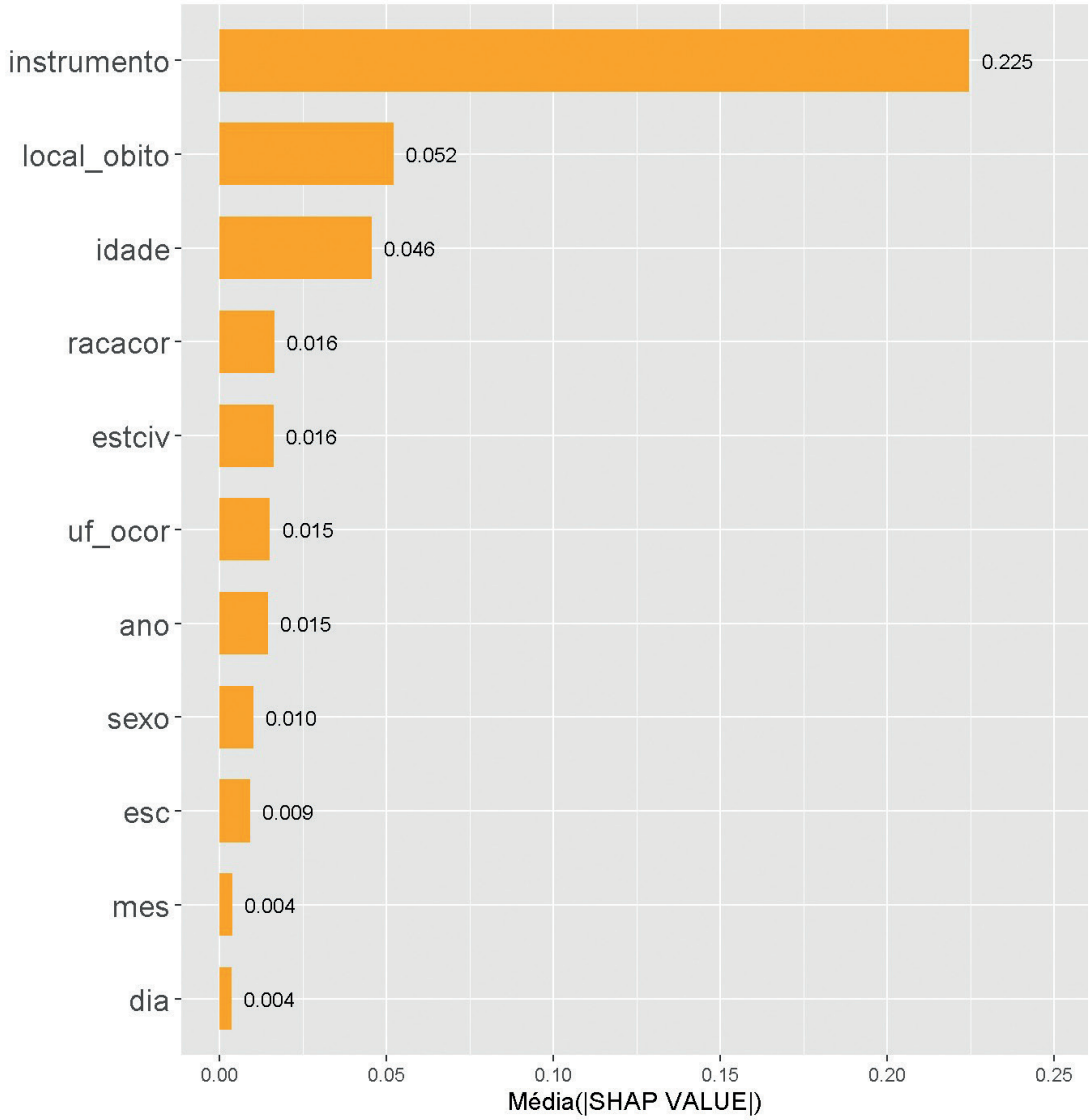
$$\emptyset_j = \sum_{S \subseteq P \setminus \{j\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \left(f_X(S\ U\ \{j\}) - f_X(S)\right)$$

Ou seja, o *SHAP* da variável *j* realiza o somatório das *S* coalizões de varáveis que não contêm a *j*-ésima variável. A contribuição marginal da *j*-ésima variável, em cada coalização, será a diferença entre a probabilidade estimada, incluindo-se a *j*-ésima variável $f_X(S\ U\ \{j\})$ e sem a *j*-ésima variável $f_X(S)$, ponderada pela probabilidade de a variável *j* contribuir na coalização *S*. Em resumo, o *SHAP* atribui importância à variável preditiva comparando a previsão do modelo com e sem variável. No entanto, como a ordem em que um modelo ajusta as variáveis pode afetar suas previsões, as comparações são realizadas em todas as possíveis ordens, para que as variáveis sejam comparadas de forma justa.

O gráfico 11 apresenta o *summary plot* da base de MVCIs, isto é, a média dos valores absolutos dos *SHAP value* de cada variável em todas as probabilidades estimadas de as MVCIs serem homicídios ocultos. O eixo y expõe, em ordem decrescente, a influência das variáveis preditivas, ou seja, o *ranking* da importância relativa de cada variável na

elaboração das probabilidades estimadas. O resultado reforça a importância da variável *instrumento* responsável pela causa básica do óbito na elaboração das classificações, com a maior média absoluta do *SHAP value* (0,225), seguido por *local do incidente* (0,052) e *idade* (0,046). A contribuição das demais variáveis parece ser negligenciável.

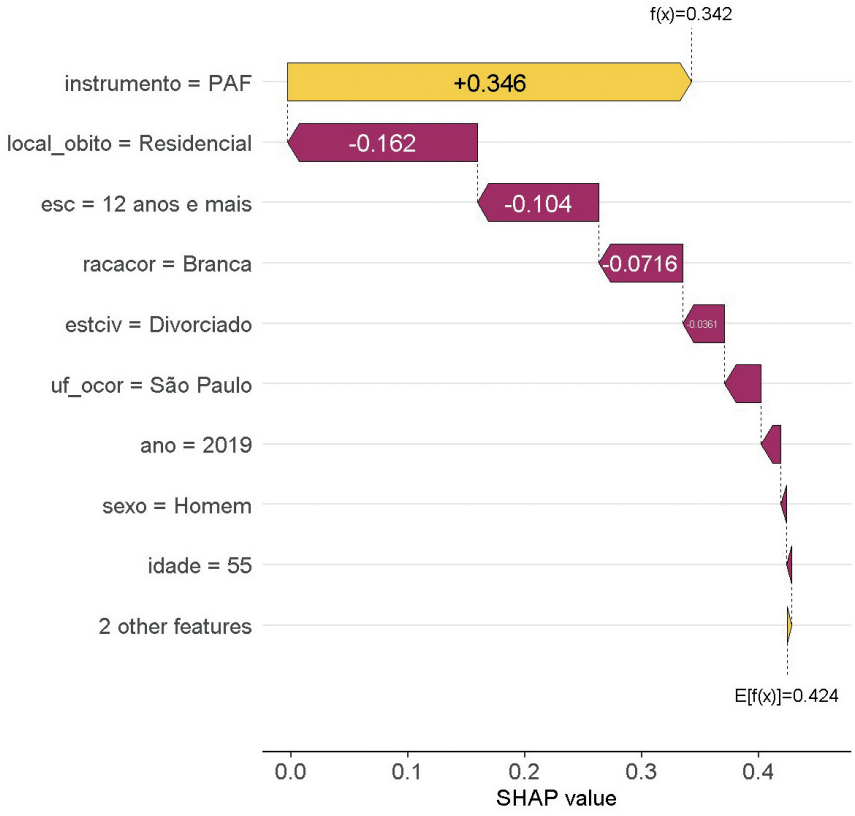
GRÁFICO 11
SHAP summary plot



Fonte: MS/SVS/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Obs.: 1. Com base no modelo *random forest*.
2. Ilustração cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

O *waterfall plot* do gráfico 12 permite observar a influência do valor das variáveis preditivas na probabilidade estimada de uma MVCI aleatoriamente selecionada ser ou não homicídio oculto. Apesar da influência positiva do instrumento PAF (+ 0.346), o valor das outras variáveis, ao se reproduzirem as relações entre as variáveis e a intencionalidade do óbito da base de óbitos conhecidos, implica influência negativa e, portanto, estima uma probabilidade de 0,342 de essa MVCI ser homicídio oculto, valor inferior ao *threshold*; portanto, essa MVCI foi classificada como acidente/suicídio.

GRÁFICO 12
***WaterFall plot*: PAF**

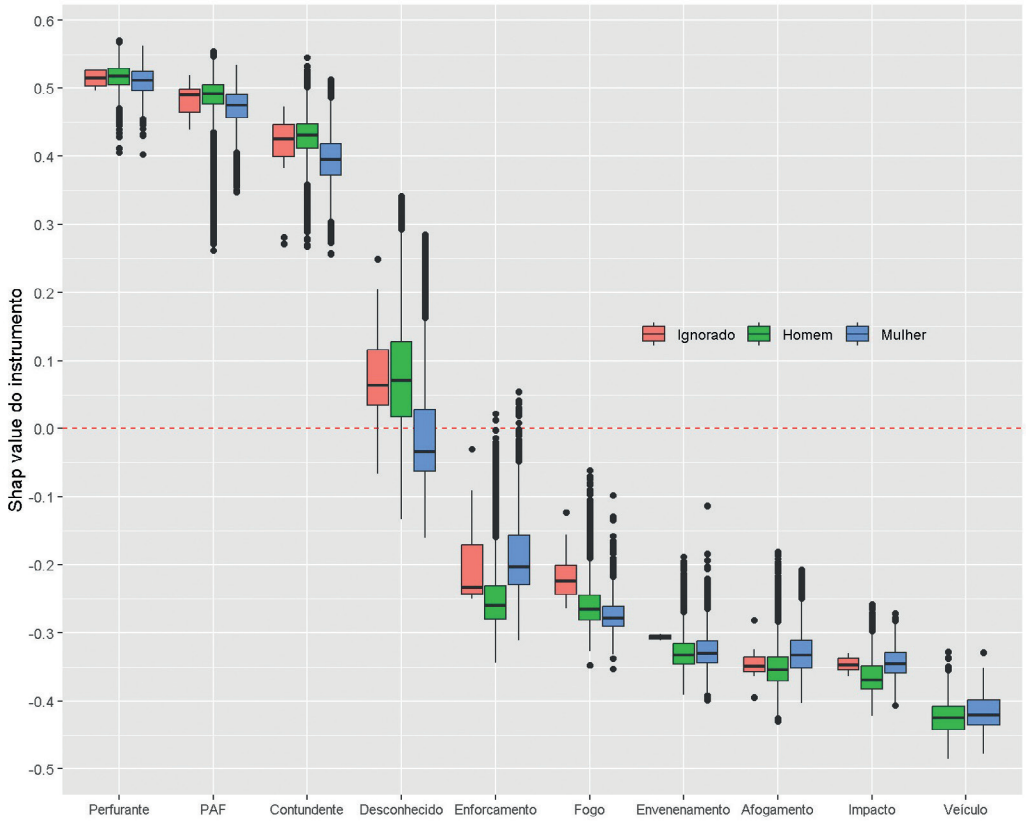


Fonte: MS/SVS/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Obs.: 1. Com base no modelo *random forest*.
2. Ilustração cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

O gráfico 13 apresenta o *boxplot* da contribuição da variável *instrumento* na probabilidade estimada de as MVCIs serem homicídios ocultos, por sexo da vítima. Assim, *instrumento perfurante*, *PAF* e *instrumento contundente* aumentam a probabilidade de a MVCI ser homicídio oculto, independente do sexo da vítima. Por outro lado, os

instrumentos *fogo*, *envenenamento*, *afogamento*, *impacto* e *veículo* reduzem a probabilidade de a MVCI ser homicídio oculto. Por fim, a contribuição do *instrumento desconhecido* parece estar condicionada a outras características da MVCI.[19]

GRÁFICO 13
***Boxplot* do *SHAP value* de instrumentos, por sexo da vítima**



Fonte: MS/SVS/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Obs.: 1. Com base no modelo *random forest*.
2. Ilustração cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

Além de mensurarmos a influência dos previsores no desempenho preditivo e na elaboração das probabilidades estimadas, precisamos entender a relação entre as variáveis preditivas e a probabilidade estimada de o óbito por causa externa de intenção indeterminada ser homicídio oculto. O método *partial dependence plots* (PDP), de Friedman (2001), apresenta a relação entre o subconjunto de variáveis previsoras e a

19. O gráfico A.2 do apêndice A apresenta o *SHAP value* de idade, por instrumento do óbito. Demais *SHAP values* estão disponíveis mediante solicitação aos autores.

probabilidade esperada, ao considerar o efeito médio de outros previsores. Na prática, o PDP da variável *j* fixada no valor *z* é a média das probabilidades estimadas ao utilizarmos o valor *z* em todas as observações da variável *j*.[20] Assim, o PDP demonstra mudanças na previsão média causada por variações no valor da variável analisada (Molnar, 2022).

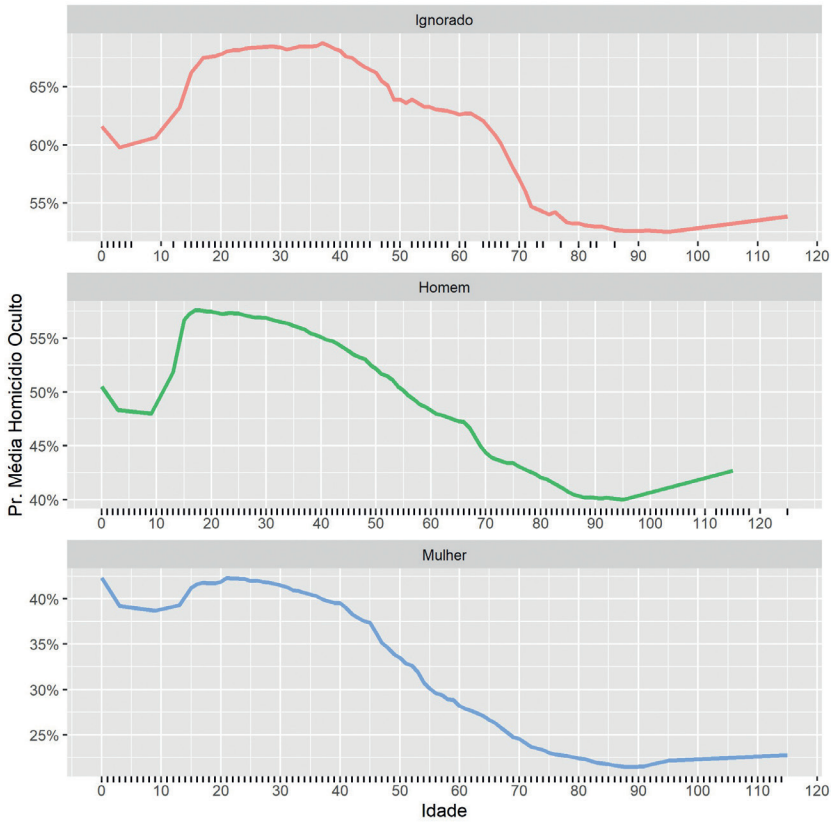
O PDP da base de MVCI, apresentado no gráfico 14, sugere não linearidade entre idade e probabilidade média de a MVCI ser homicídio oculto, independentemente do sexo analisado. Assim, a maior chance de uma mulher ser vítima de homicídio oculto (42,3%) acontece logo ao nascer, isto é, enquanto a idade é inferior a 1 ano. Em seguida, entre 15 e 30 anos, ocorrem platô (próximo à probabilidade máxima, reflexo da elevada ocorrência de vítimas jovens) e gradual redução a partir dos 30 anos, acentuada após os 40 anos. Em vítimas do sexo feminino, a probabilidade média é sempre inferior a 50%, reflexo da menor ocorrência de homicídios entre mulheres.

No caso de vítimas do sexo masculino, existe acentuado aumento na probabilidade média estimada de homicídio oculto entre 10 e 19 anos, na transição de idades entre criança e jovem, reflexo do elevado número de homicídios nessa faixa etária. Homens com 17 anos têm a maior probabilidade média (57,5%) de vitimização por homicídio oculto. Após a máxima da probabilidade média, ocorre monótona redução na probabilidade de vitimização; ainda assim, em homens com idade entre 12 e 55 anos, a probabilidade média é sempre superior a 50%.

O resultado de vítimas com sexo ignorado é semelhante àquele de vítimas masculinas, isto é, probabilidade crescente entre jovens e redução monótona após probabilidade máxima. Entretanto, a interpretação do resultado de vítimas com sexo ignorado requer ressalvas, por causa do reduzido número de observações, especialmente nos valores extremos de idade.

20. O PDP do modelo f() e variável previsora x^j fixada no valor *z* é definido por: $g^j_{PDP}(z) = E_{X^{-j}}\{f(X^{j|=z})\}$. Não conhecemos a distribuição marginal de X^{-j}. Entretanto, é possível estimar a distribuição marginal de X^{-j} através das *n* observações da base investigada. Assim, o valor estimado do PDP é $\hat{g}^j_{PDP}(z) = \frac{1}{n}\sum_{i=1}^{n} f(\underline{x}_i^{j|=z})$ – Biecek e Burzykowski (2021).

GRÁFICO 14
***Partial dependence plot* de idade, por sexo**
(Em %)

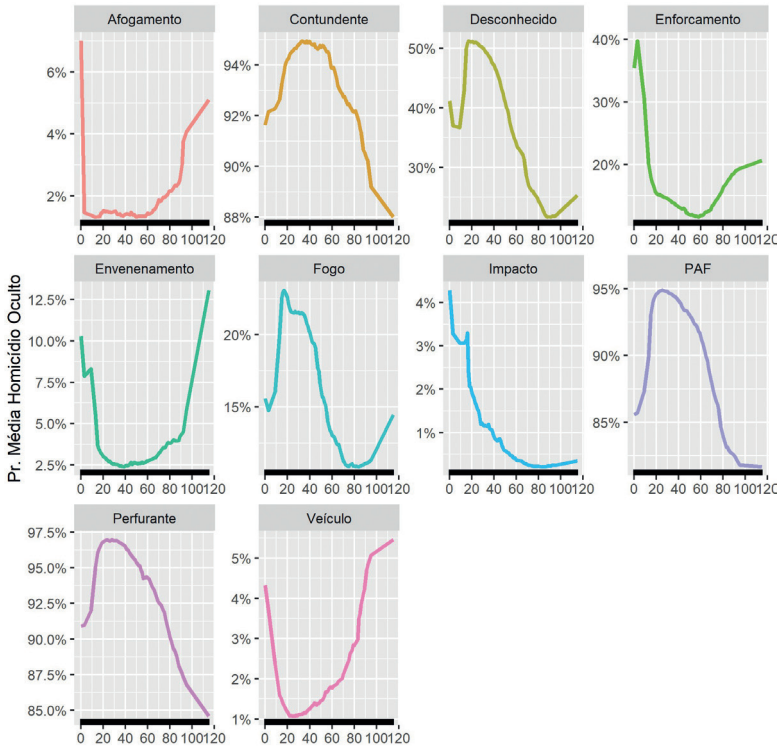


Fonte: MS/SVS/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Obs.: 1. Com base no modelo *random forest*.
2. Traços no eixo x indicam a distribuição de idade.
3. Ilustração cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

No gráfico 15, ao se analisar o PDP entre idade e instrumentos, a evidência atribui maior probabilidade média de homicídio oculto aos instrumentos de elevada representatividade em homicídios. Assim, os instrumentos *contundente*, *perfurante* e *PAF* registram elevada probabilidade média de vitimização por homicídio oculto, não linearidade na relação com idade, e probabilidades máximas entre vítimas jovens. Em homicídios ocultos causados por fogo, impacto e instrumento desconhecido, a relação não linear entre a idade e o instrumento reproduz a incidência desses em homicídio. Nos casos de afogamento, enforcamento e envenenamento, as probabilidades máximas, apesar de reduzidas, estão nas extremidades de idades, reflexo dos casos de infanticídio e gerontocídio. Por sua vez, no instrumento *veículo*, o formato da curva reproduz a elevada representatividade

de jovens em acidentes (35,3%), e o reduzido patamar da probabilidade média reflete a baixa ocorrência de homicídios causados por veículos (0,1%).

GRÁFICO 15
***Partial dependence plot* de idade, por instrumento do óbito**
(Em %)



Fonte: MS/SVS/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Obs.: 1. Com base no modelo *andom forest*.
2. Traços no eixo x indicam a distribuição de idade.
3. Ilustração cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

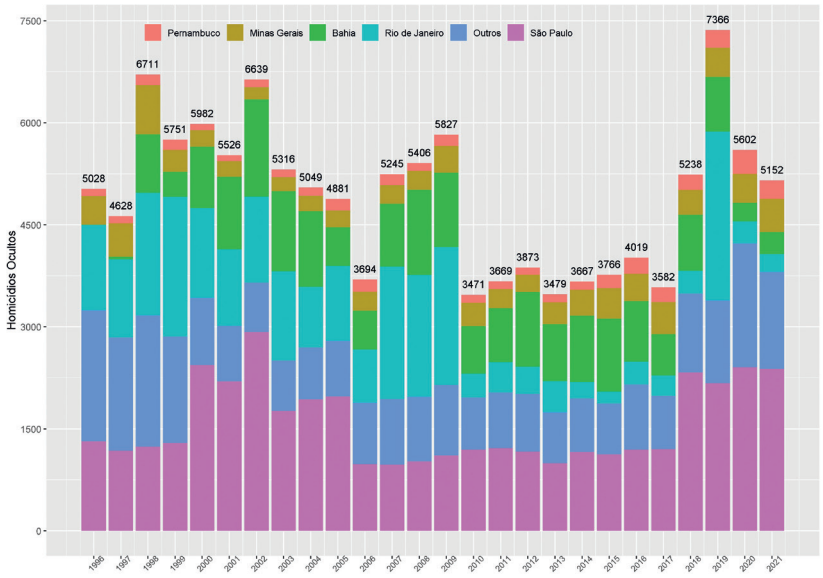
Portanto, os resultados apresentados nesta seção indicam adequada capacidade de generalização do modelo utilizado (*random forest*), sugerido no bom desempenho preditivo em observações inéditas da base de teste, e a importância do instrumento responsável pela causa básica do óbito na classificação de novas observações e no desempenho preditivo do modelo. A classificação do óbito por causa externa de intenção indeterminada como homicídio oculto sofre influência relevante dos instrumentos *PAF*, *impacto* e *contundente* em relação não linear entre os instrumentos e a idade da vítima, com valor máximo da probabilidade média de homicídio oculto entre jovens.

4 HOMICÍDIOS OCULTOS NO BRASIL

Estimamos que, durante o período 1996-2021, o sistema de saúde não identificou 128.567 homicídios. Portanto, os achados sugerem que 43,6% das MVCIs registradas no Brasil são, na realidade, homicídios ocultos. Esse valor revela-se superior em 10 p.p. ao encontrado por Andreev *et al*. (2015) em estudo similar que investigou as MVCIs na Rússia. Na média anual, o Estado foi incapaz de identificar 4.492 homicídios ao ano, valor próximo à mediana de homicídios da Bahia. Em outras palavras, em cada ano do período analisado, não são registrados, em média, um número de homicídios semelhante àquele observado na UF com o terceiro maior número absoluto de homicídios. O achado sugere que, ao menos no caso brasileiro, ao contrário do preconizado em Pinotti (2020), a extensão da subenumeração de homicídios não é negligenciável.

O gráfico 16 ilustra a trajetória temporal dos homicídios ocultos no Brasil, entre 1996 e 2021, para as UFs que apresentam os cinco maiores números absolutos de homicídios registrados e também de homicídios ocultos, a cada ano. São Paulo, Rio de Janeiro, Bahia, Minas Gerais e Pernambuco respondem conjuntamente por 78,2% do total de homicídios ocultos no país. É notável que essas UFs permanecem, ao longo de todo o período, entre aquelas com os cinco maiores números absolutos de homicídios ocultos.

GRÁFICO 16
Decomposição temporal dos homicídios ocultos, por UF (1996-2021)



Fonte: MS/SVS/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.

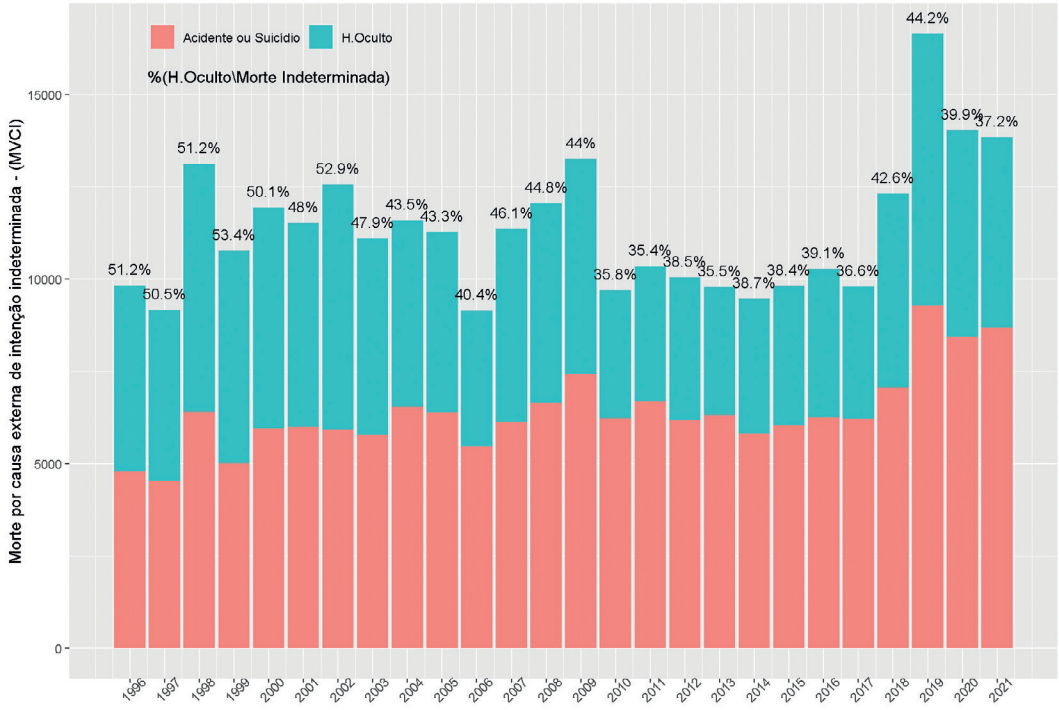
Obs.: Ilustração cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

O gráfico 17 apresenta a trajetória da proporção de homicídios ocultos incorretamente classificados na categoria MVCI, expondo três padrões de subenumeração. No intervalo entre 1996 e 2009, a média das proporções de 47,6% foi a maior em todo o período. Nesse contexto de elevada incerteza, o ano de 1999 registrou a maior proporção da série, com 53,4%. A partir de 2010, e durante os sete anos subsequentes, verifica-se retração no valor absoluto de MVCI, ao passo que a proporção média dos homicídios ocultos nas MVCIs atinge 37,1%, menor média da série, com valor mínimo de 35,4%. Ainda assim, nesses anos de menor subenumeração, ao menos um terço das MVCIs são homicídios ocultos. No quadriênio final, a incerteza sobre intencionalidade dos óbitos volta a aumentar e registra-se, em 2019, o maior número absoluto de MVCIs. Entretanto, não ocorre expansão de mesma magnitude nos homicídios ocultos, embora a média anual tenha aumentado para 40,9% nesse mesmo quadriênio.

GRÁFICO 17

Percentual estimado de homicídios ocultos e de acidentes ou suicídios ocultos em relação ao total de MVCIs (1996-2021)

(Em %)



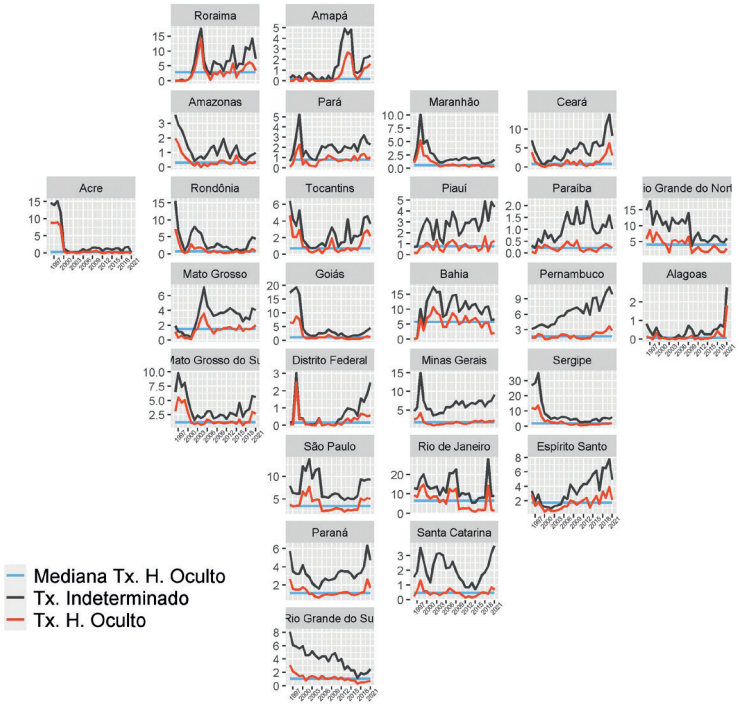
Fonte: MS/SVS/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Obs.: Ilustração cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

Ao evitarem o viés estatístico decorrente da subnotificação de indicadores criminais, alguns estudos assumem a premissa de que a proporção de subnotificação permanece constante ao longo do tempo, nas áreas sob investigação (por exemplo, Bianchi, Buonanno e Pinotti, 2012). A evidência apresentada aponta substancial variabilidade na proporção de homicídios subenumerados, contrariando, ao menos no caso dos homicídios brasileiros, a hipótese geralmente adotada na literatura.

O gráfico 18 apresenta a evolução da taxa de homicídios ocultos por 100 mil habitantes nas UFs e sugere elevada dispersão na ocorrência de homicídios ocultos entre UFs e, em alguns casos, na própria UF. Rio de Janeiro, Bahia e Rio Grande do Norte apresentam, relativamente às outras UFs, elevadas taxas de homicídio oculto ao longo de praticamente todo o período, com mediana da taxa de homicídio oculto próxima a 5. Nesse grupo, o Rio Grande do Norte, a partir de 2010, conseguiu reduzir o erro de identificação sobre as intencionalidades das mortes, registrando uma média anual de 2,3 entre 2010 e 2021, o que é 3,1 inferior à média anual do período anterior. Por sua vez, a Bahia mantém taxas de homicídio oculto elevadas ao longo de todo o período. O Rio de Janeiro, após melhora iniciada em 2010, registrou, em 2019, a segunda maior taxa de homicídios ocultos do país, 14,4.

No gráfico, podemos ainda perceber que Alagoas, Distrito Federal e Amapá possuem as menores medianas da taxa de homicídio. As UFs que conseguiram reduzir as taxas de homicídios ocultos, ao longo do período analisado, foram Acre, Amazonas, Maranhão, Mato Grosso do Sul, Goiás e Rio Grande do Sul. Por fim, Minas Gerais e São Paulo, após período de taxas elevadas, se estabilizaram à taxa de aproximadamente 4 homicídios ocultos por 100 mil habitantes, embora São Paulo apresente expressiva degradação na qualidade dos dados ao final do período analisado. O crescimento lento do denominador populacional, associado à elevada variabilidade das taxas, indica que, de modo semelhante à série brasileira, a proporção de homicídios ocultos nas UFs não se mantém constante ao longo do tempo.

GRÁFICO 18
Taxa de homicídios ocultos, por UFs (1996-2021)

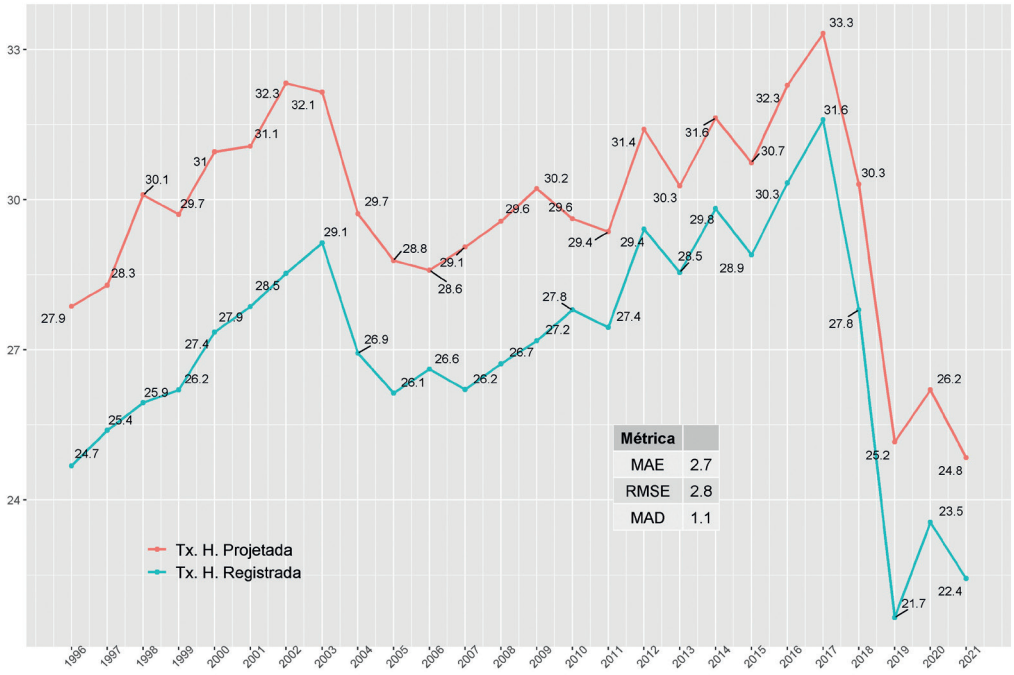


Fonte: MS/SVS/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Obs.: Ilustração cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

4.1 Número projetado de homicídios

Apesar do elevado número de homicídios ocultos, a soma destes aos homicídios registrados, isto é, os homicídios projetados, parece não provocar quebra da série temporal de homicídios no Brasil. De acordo com o gráfico 19, a diferença entre a taxa de homicídios registrados e projetados parece restrita ao nível das séries, sem impacto na tendência. Todavia, a comparação das séries, utilizando-se métricas preditivas, denota diferença relevante – por exemplo, 2,8 unidades no caso da raiz do erro quadrático médio (RMSE). O elevado valor de erro sugere que a contabilização dos homicídios ocultos traz novo sentido à análise sobre a taxa brasileira de homicídios. Por exemplo, enquanto a taxa de crescimento anual de 2019 indica redução de 22,1% na taxa de homicídios registrados, no caso da taxa de homicídios projetados, esse indicador mostra redução de 17,0%, uma diferença de 5 p.p., resultado relevante para avaliações de políticas públicas.

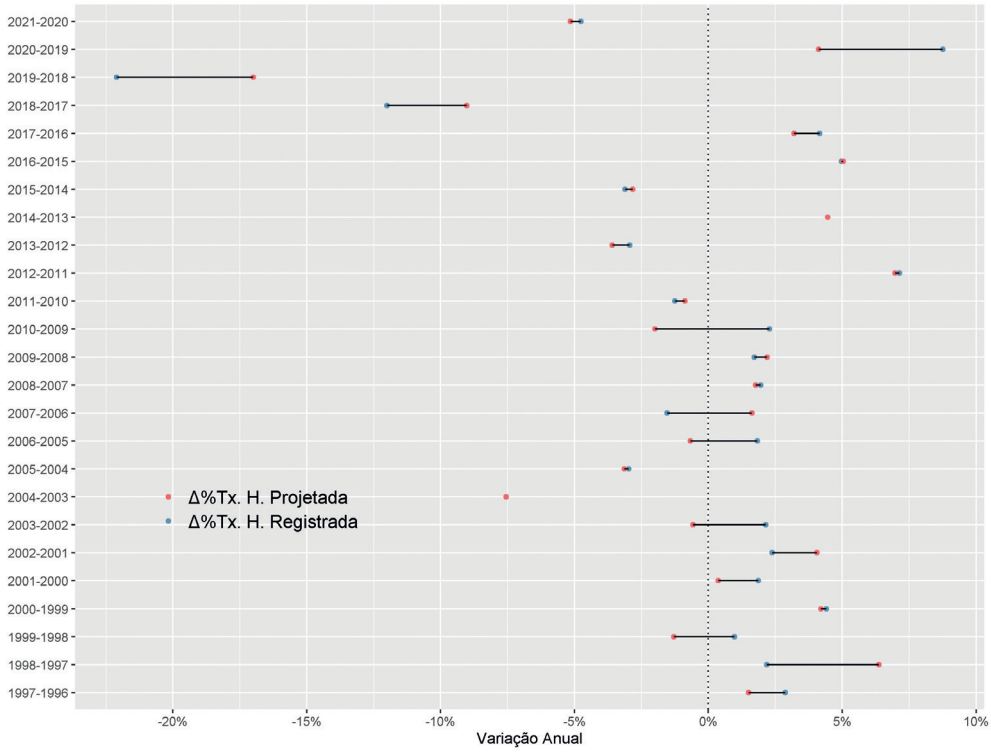
GRÁFICO 19
Taxas de homicídios registrados *versus* taxas de homicídios projetados – Brasil (1996-2021)
(Em %)



Fonte: MS/SVS/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Obs.: Ilustração cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

A diferença entre a taxa de crescimento anual da taxa de homicídios registrados e projetados sintetiza o impacto da inclusão dos homicídios ocultos. De acordo com o gráfico 20, as maiores distâncias entre as taxas de crescimento anual ocorrem no início e no final da série analisada, períodos de elevado número de homicídios ocultos, achado que sozinho sugere nova interpretação da dinâmica criminal. Adicionalmente, em períodos intermediários, verifica-se inversão na direção da variação anual em cinco ocasiões, o que altera diagnósticos de segurança pública e revela a importância de se questionar a hipótese de subnotificação reduzida em homicídios. Como resultado das diferentes variações anuais, em média, a taxa de homicídios projetada supera a taxa registrada em 10,0% e, no período analisado, o acumulado de homicídios projetados superou os registrados em 9,6%, totalizando 1.454.544 homicídios.

GRÁFICO 20
Diferencial do crescimento percentual anual entre as taxas de homicídios registrados e projetados – Brasil (1996-2021)
(Em %)



Fonte: MS/SVS/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Obs.: Ilustração cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

O impacto, ao longo do tempo, dos homicídios ocultos nas taxas estaduais de homicídios aparece sintetizado no gráfico 21, que apresenta as taxas por 100 mil habitantes de homicídios registrados e projetados entre as dez UFs mais violentas de cada ano.[21] Incluir homicídios ocultos no cálculo das taxas estaduais de homicídio acarreta, em todos os anos, troca de posições entre as UFs mais violentas e, em oito anos, ocorre mudança no posto de UF mais violenta.

Em particular, devemos notar que, em 2019, somente após contabilizar os homicídios ocultos, o Rio de Janeiro aparece entre as dez UFs mais violentas. Nesse ano,

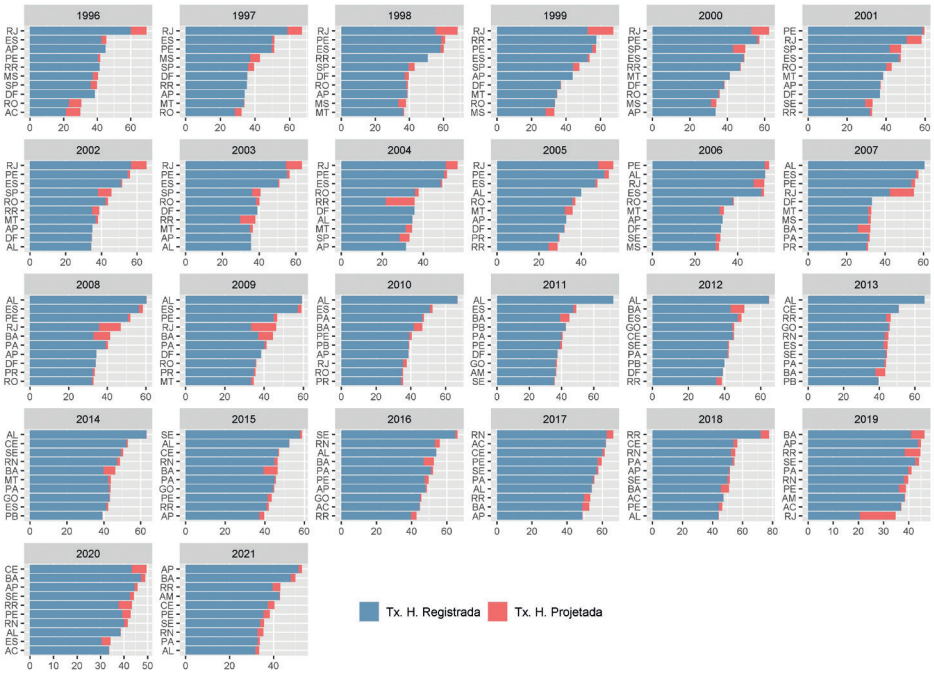
21. O impacto dos homicídios ocultos, em valores absolutos, está apresentado na figura A.1 do apêndice A. O resultado é qualitativamente igual ao reportado no caso das taxas por 100 mil habitantes.

são identificados no Rio de Janeiro 2.480 homicídios ocultos, valor correspondente a 70,0% dos homicídios registrados. Em quatro ocasiões, Roraima também aparece entre as UFs mais violentas somente após a inclusão dos homicídios ocultos. Esses achados evidenciam o imperativo de se incluírem os homicídios ocultos em investigações sobre criminalidade violenta. A ausência destes impossibilita diagnósticos capazes de focalizar intervenções nas áreas mais críticas ou caracterizar as populações mais vitimizadas, impedindo a orientação de políticas para redução da criminalidade violenta.

GRÁFICO 21

Dez maiores taxas de homicídios de UFs, por ano (1996-2021)

(Em %)



Fonte: MS/SVS/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Obs.: Ilustração cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

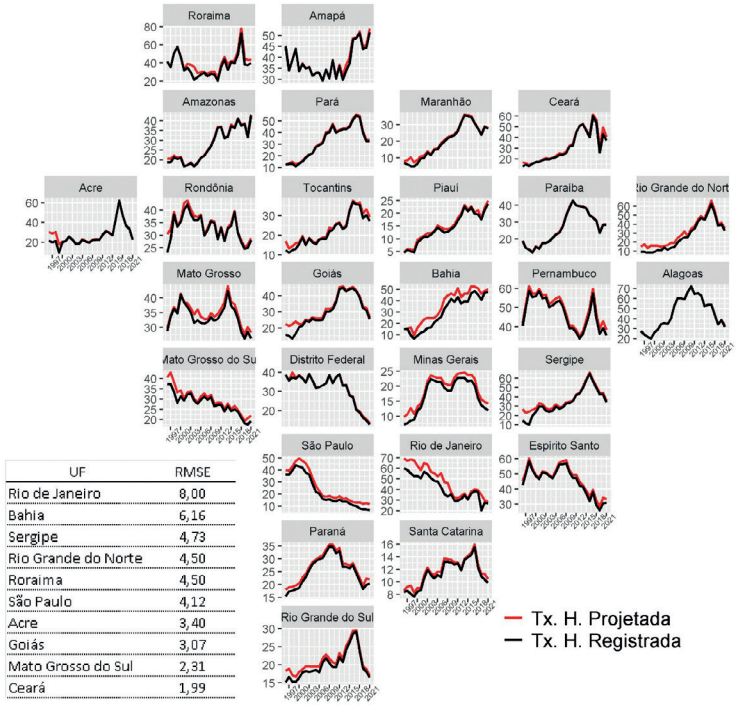
O gráfico 22 apresenta o mapa do Brasil com as séries temporais das taxas estaduais de homicídio por 100 mil habitantes registradas e projetadas. De modo similar ao observado na série brasileira, as taxas de homicídio projetadas seguem as taxas de homicídio registradas, apesar das significativas diferenças em nível. Exercícios de similaridade entre as taxas encontram valores elevados da métrica RMSE e indicam as UFs mais afetadas, conforme tabela apresentada no canto inferior do gráfico. Seis UFs (Rio de Janeiro, Bahia, Sergipe, Rio Grande do Norte, Roraima e São Paulo) aparecem

com RMSE superior a 4, resultado relevante, dadas as médias estaduais de homicídio registrado. Por exemplo, em São Paulo, ao longo de todo o período, a taxa projetada é em média 17,7% superior à registrada, ou 69,1%, considerando-se somente o quadriênio final.

Esse novo retrato da criminalidade violenta produz efeitos em diagnósticos e avaliações de políticas de segurança. Por exemplo, no Rio de Janeiro, a redução nos homicídios registrados durante o período de implementação acelerada das Unidades de Polícia Pacificadora (UPPs),[22] isto é, entre 2009 e 2013, de -6,9% na taxa de homicídios registrados, torna-se redução de -26,3% na taxa de homicídios projetados, efeito causado por elevada subnotificação em 2009. Ainda no Rio de Janeiro, em 2019, período de elevada incerteza sobre a intencionalidade dos óbitos na UF, a taxa de crescimento anual de homicídios registrados e projetados apresenta discrepância singular em relação às demais UFs. Enquanto a taxa de homicídios registrada apresentou redução de -45,6%, a taxa de homicídios projetada sugere retração de -12,1%.

GRÁFICO 22

Taxas de homicídios observados e projetados, por UFs (1996-2021)



UF	RMSE
Rio de Janeiro	8,00
Bahia	6,16
Sergipe	4,73
Rio Grande do Norte	4,50
Roraima	4,50
São Paulo	4,12
Acre	3,40
Goiás	3,07
Mato Grosso do Sul	2,31
Ceará	1,99

Fonte: MS/SVS/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Obs.: Ilustração cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

22. Para consultar detalhes, ver Montes e Lins (2018).

5 CONCLUSÃO

O SIM/MS é referência sobre mortes violentas e, em particular, homicídios no Brasil, sendo fonte de informação para diversos estudos sobre criminalidade (Cerqueira e Soares, 2016).

Entretanto, a ocorrência de subenumeração de óbitos no SIM/MS (Soares Filho, Cortez-Escalante e França, 2016), bem como o registro do incidente em categorias mal definidas, conspira contra a qualidade dos dados e contribui para a subnotificação dos homicídios. Tal fenômeno não é exclusividade do Brasil. As evidências encontradas nas literaturas internacional (Andreev *et al*., 2015; Santoro, 2020) e nacional (Melo *et al*., 2014) sugerem que uma parcela dos óbitos por MVCI são, na realidade, homicídios ocultos.

No Brasil, entre 1996 e 2021, ocorreram 3.396.010 mortes violentas,[23] sendo que a causa básica do óbito não foi definida em 294.752, ou em 8,7% do total. O propósito deste trabalho foi estimar qual parcela dessas MVCIs se constituía, na verdade, de homicídios ocultos.

A metodologia adotada baseou-se em um conjunto de modelos de aprendizado supervisionado (*machine learning*), em que o padrão probabilístico de características pessoais e situacionais para cada tipo de evento (se homicídio ou se suicídio/acidente) foi aprendido. Com base na escolha do modelo com melhor capacidade de generalização, classificamos as MVCIs como homicídios, ou como acidentes ou suicídios.

Nossas estimativas indicaram ter havido, no período 1996-2021, 128.567 homicídios ocultos, o que representou 43,6% do total de MVCIs. Em média, o número de homicídios ocultos ao ano foi de 4.492. Esse número corresponde à média anual de homicídios que ocorre no estado de São Paulo, ou à queda sem sobreviventes de 33 Boeings 787 lotados, em tragédias totalmente invisibilizadas.

Considerando-se o número de homicídios projetados como a soma dos homicídios registrados e homicídios ocultos, em média, a taxa de homicídios projetada por 100 mil habitantes supera a taxa registrada em 8,3%. Nesse período, ao passo que o país registrou oficialmente 1.325.977 homicídios, nossas estimativas indicam um número de 1.454.544 casos.

Enquanto a série temporal de homicídios no Brasil não sofreu transformação significativa, a quantidade de homicídios ocultos foi grande o suficiente para alterar diagnósticos

23. Como apontado anteriormente, foram excluídos os óbitos por sequelas médicas, privações e acidentes naturais.

e avaliações de políticas de públicas, bem como invalidar procedimentos econométricos usualmente adotados na correção de subnotificação. Por exemplo, variações na distribuição temporal dos homicídios ocultos entre UFs e frequência da subenumeração invalidam estratégias tradicionalmente utilizadas, ao serem corrigidas subnotificações no contexto de painel de dados (Bianchi, Buonanno e Pinotti, 2012). A evidência de haver um número considerável de homicídios ocultos relativiza a redução de homicídios em algumas UFs, troca de posições entre UFs mais violentas e inversão no sentido da taxa de variação anual dos homicídios brasileiros, alterando o entendimento da dinâmica criminal.

A análise subnacional acerca do expressivo número de MVCIs e de homicídios ocultos indica, em última instância, não ser um problema generalizado no país, mas estar concentrado em quatro UFs. Com efeito, São Paulo, Rio de Janeiro, Bahia e Minas Gerais são responsáveis por 72,5% dos homicídios ocultos no Brasil.

O método proposto, no entanto, não esgota todas as possibilidades de identificação de homicídios ocultos, e não deve ser entendido como substituto da necessidade de se aprimorar a qualidade das informações do SIM/MS. Em particular, uma questão para a qual a academia ainda não tem uma resposta diz respeito ao número de homicídios não registrados que ocorrem no país, quando não há sequer uma declaração de óbito, como no exemplo dos inúmeros casos revelados no Rio de Janeiro, em que milicianos "somem" com o corpo de suas vítimas.

REFERÊNCIAS

AAS, K.; JULLUM, M.; LOLAND, A. Explaining individual predictions when features are dependent: more accurate approximations to Shapley values. **Artificial Intelligence**, v. 298, p. 1-24, set. 2021.

ANDREEV, E. *et al*. A method for reclassifying cause of death in cases categorized as "event of undetermined intent". **Population Health Metrics**, v. 13, n. 23, p. 1-25, 2 dez. 2015.

BENAVOLI, A. *et al*. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. **Journal of Machine Learning Research**, v. 18, n. 77, p. 1-36, 2017.

BIANCHI, M.; BUONANNO, P.; PINOTTI, P. Do immigrants cause crime? **Journal of the European Economic Association**, v. 10, n. 6, p. 1318-1347, dez. 2012.

BIECEK, P. Dalex: explainers for complex predictive models in R. **Journal of Machine Learning Research**, v. 19, p. 1-5, 2018.

BIECEK, P.; BURZYKOWSKI, T. Partial-dependence profiles. *In*: BIECEK, P.; BURZYKOWSKI, T. **Explanatory model analysis**: explore, explain, and examine predictive models. 1. ed. Nova York: CRC Press, 2021. p. 209-226.

BRASIL. Lei nº 6.015, de 31 de dezembro de 1973. Dispõe sobre os registros públicos, e dá outras providências. **Diário Oficial da União**, Brasília, 31 dez. 1973. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l6015compilada.htm.

BRASIL. Ministério da Saúde. Portaria nº 116, de 11 de fevereiro de 2009. Regulamenta a coleta de dados, fluxo e periodicidade de envio das informações sobre óbitos e nascidos vivos para os Sistemas de Informações em Saúde sob gestão da Secretaria de Vigilância em Saúde. **Diário Oficial da União**, Brasília, 11 fev. 2009. Disponível em: http://bvsms.saude.gov.br/bvs/saudelegis/svs/2009/prt0116_11_02_2009.html.

BRASIL. **Declaração de óbito**: manual de instruções para preenchimento. Brasília: Ministério da Saúde, 2022.

BRASIL; CFM – CONSELHO FEDERAL DE MEDICINA; CBCD – CENTRO BRASILEIRO DE CLASSIFICAÇÃO DE DOENÇAS. **A declaração de óbito**: documento necessário e importante. 3. ed. Brasília: Ministério da Saúde, 2009.

BRODLEY, C. E.; FRIEDL, M. A. Identifying mislabeled training data. **Journal of Artificial Intelligence Research**, v. 11, p. 131-167, 1º ago. 1999.

CAVALINI, L. T.; LEON, A. C. M. P. de. Correção de sub-registros de óbitos e proporção de internações por causas mal definidas. **Revista de Saúde Pública**, v. 41, n. 1, p. 85-93, fev. 2007.

CERQUEIRA, D. Mortes violentas não esclarecidas e impunidade no Rio de Janeiro. **Economia Aplicada**, v. 16, n. 2, p. 201-235, 2012.

CERQUEIRA, D. **Mapa dos homicídios ocultos no Brasil**. Brasília: Ipea, jul. 2013. (Texto para Discussão, n. 1848).

CERQUEIRA, D.; SOARES, R. R. The welfare cost of homicides in Brazil: accounting for heterogeneity in the willingness to pay for mortality reductions. **Health Economics**, v. 25, n. 3, p. 259-276, mar. 2016.

CFM – CONSELHO FEDERAL DE MEDICINA. Resolução CFM nº 1.779/05. Regulamenta a responsabilidade médica no fornecimento da declaração de óbito. **Diário Oficial da União**, Brasília, p. 121, 11 nov. 2005. Seção 1. Disponível em: https://sistemas.cfm.org.br/normas/visualizar/resolucoes/BR/2005/1779.

CHAWLA, N. V. *et al*. Smote: synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321-357, 1º jun. 2002.

DIÓGENES, V. H. D. *et al*. Differentials in death count records by databases in Brazil in 2010. **Revista de Saúde Pública**, v. 56, n. 92, p. 1-11, 24 out. 2022.

FISHER, A.; RUDIN, C.; DOMINICI, F. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. **Journal of Machine Learning Research**, v. 20, p. 1-81, 2019.

FRIAS, P. G. de. *et al*. Utilização das informações vitais para a estimação de indicadores de mortalidade no Brasil: da busca ativa de eventos ao desenvolvimento de métodos. **Cadernos de Saúde Pública**, v. 33, n. 3, p. 1-13, 2017.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. **The Annals of Statistics**, v. 29, n. 5, p. 1189-1232, 1º out. 2001.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. Model assessment and selection. *In*: HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning**. 2. ed. Nova York: Springer, 2009. p. 219-259.

JAMES, G. *et al*. Resampling methods. *In*: JAMES, G. *et al*. **An introduction to statistical learning**: with applications in R. 2. ed. Nova York: Springer, 2021. p. 197-223.

KUHN, M.; JOHNSON, K. Over-fitting and model tuning. *In*: KUHN, M.; JOHNSON, K. **Applied predictive modeling**. Nova York: Springer, 2013. p. 61-92.

LOPES, A. S. *et al*. Melhoria da qualidade do registro da causa básica de morte por causas externas a partir do relacionamento de dados dos setores saúde, segurança pública e imprensa, no estado do Rio de Janeiro, 2014. **Epidemiologia e Serviços de Saúde**, Brasília, v. 27, n. 4, p. 1-10, 2018.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. *In*: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 31., 2017, Long Beach, Califórnia. **Anais**... 2017.

MELO, C. M. de. *et al*. Qualidade da informação sobre óbitos por causas externas em município de médio porte em Minas Gerais, Brasil. **Cadernos de Saúde Pública**, Rio de Janeiro, v. 30, n. 9, p. 1999-2004, set. 2014.

MOLNAR, C. **Interpretable machine learning**: a guide for making black box models explainable. 2. ed. [s.l.]: [s.n.], 2022.

MONTES, G. C.; LINS, G. O. Deterrence effects, socio-economic development, police revenge and homicides in Rio de Janeiro. **International Journal of Social Economics**, v. 45, n. 10, p. 1406-1423, set. 2018.

MURPHY, K. P. **Probabilistic machine learning**: an introduction. 1. ed. Cambridge, Estados Unidos: MIT Press, 2020.

NASCIMENTO, C. F. do. *et al*. Cause-specific mortality prediction in older residents of São Paulo, Brazil: a machine learning approach. **Age and Ageing**, v. 50, n. 5, p. 1692-1698, 11 set. 2021.

OHBERG, A.; LONNQVIST, J. Suicides hidden among undetermined deaths. **Acta Psychiatrica Scandinavica**, v. 98, n. 3, p. 214-218, 1998.

OPEN SOCIETY FOUDATIONS *et al*. Protocolo de Bogotá: sobre a qualidade dos dados de homicídio na América Latina e Caribe. *In*: CONFERÊNCIA SOBRE QUALIDADE DE DADOS DE HOMICÍDIOS NA AMÉRICA LATINA E NO CARIBE, 2015, Bogotá, Colômbia. **Anais**... 2015. Disponível em: uerj.org/protocolo-de-bogota-sobre-calidad-de-los-datos-de-homicidio-en-america-latina-y-el-caribe/.

PINOTTI, P. The credibility revolution in the empirical analysis of crime. **Italian Economic Journal**, v. 6, n. 2, p. 207-220, 23 jul. 2020.

SANTORO, A. Recálculo de las tendencias de mortalidad por accidentes, suicidios y homicidios en Argentina, 1997-2018. **Revista Panamericana de Salud Publica**, v. 44, p. 1-6, 2020.

SHAPLEY, L. S. A value for n-person games. *In*: KUHN, H. W.; TUCKER, A. W. (Ed.). **Contributions to the theory of games (AM-28)**. Princeton: Princeton University Press, 1953. v. 2, p. 307-318.

SOARES FILHO, A. M.; CORTEZ-ESCALANTE, J. J.; FRANÇA, E. Review of deaths correction methods and quality dimensions of the underlying cause for accidents and violence in Brazil. **Ciência e Saúde Coletiva**, v. 21, n. 12, p. 3803-3818, 2016.

TABARROK, A.; HEATON, P.; HELLAND, E. The measure of vice and sin: a review of the uses, limitations and implications of crime data. *In*: BENSON, B. L.; ZIMMERMAN, P. R. (Ed.). **Handbook on the economics of crime**. Massachusetts: Edward Elgar Publishing, 2010. p. 53-81.

VÄRNIK, P. *et al*. Massive increase in injury deaths of undetermined intent in ex-USSR Baltic and Slavic countries: hidden suicides? **Scandinavian Journal of Public Health**, v. 38, n. 4, p. 395-403, 18 jun. 2010.

BIBLIOGRAFIA COMPLEMENTAR

BIAU, G. Analysis of a random forests model. **Journal of Machine Learning Research**, v. 13, p. 1063-1095, 2012.

BJÖRKENSTAM, C. *et al*. Suicide or undetermined intent? A register-based study of signs of misclassification. **Population Health Metrics**, v. 12, p. 1-11, 17 dez. 2014.

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

BREIMAN, L. *et al*. **Classification and regression trees**. Londres: Routledge, 1984.

DIRK, R. Comparação entre os registros de ocorrência (PCERJ) e as declarações de óbitos (SVS-SES/RJ). **Cadernos de Segurança Pública**, v. 9, n. 8, p. 75-83, jul. 2017.

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861-874, jun. 2006.

FERNÁNDEZ, A. *et al*. Performance measures. *In*: FERNÁNDEZ, A. *et al*. **Learning from imbalanced data sets**. Cham: Springer, 2018. p. 47-61.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. Random forests. *In*: HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning**. 2. ed. Nova York: Springer, 2009. p. 587-604.

KUHN, M.; JOHNSON, K. Measuring performance in classification models. *In*: KUHN, M.; JOHNSON, K. **Applied predictive modeling**. Nova York: Springer, 2013. p. 247-273.

MONTEIRO, J.; CABALLERO, B. Crime e violência. *In*: SHIKIDA, C. D.; MONASTERIO, L.; NERY, P. F. (Ed.). **Guia brasileiro de análise de dados**: armadilhas e soluções. 1. ed. Brasília: Enap, 2021. p. 126-169.

SOARES, L. E. **Meu casaco de general**. Rio de Janeiro: Companhia das Letras, 2000.

WHO – WORLD HEALTH ORGANIZATION. **International statistical classification of diseases and related health problems**. 5. ed. Genebra: WHO, 2016.

ZILLI, L. F. Mensurando a violência e o crime: potencialidades, vulnerabilidades e implicações para políticas de segurança pública. **Revista Brasileira de Segurança Pública**, São Paulo, v. 12, n. 1, p. 30-48, fev.-mar. 2018.

ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. **Journal of the Royal Statistical Society Series B**, v. 67, n. 2, p. 301-320, abr. 2005.

APÊNDICE A

TABELA A.1
Instrumentos e causas do óbito (1996-2021)

Instrumento	Homicídio				Não homicídio				Indeterminado		Total
	Agressão		Intervenção legal		Acidente		Suicídio				
	Quantidade	(%)	Quantidade	(%)	Quantidade	(%)	Quantidade	(%)	Quantidade	(%)	
Afogamento	1.526	0,9	-	0,0	149.446	88,4	2.874	1,7	15.268	9,0	**169.114**
Contundente	87.630	71,3	136	0,1	401	0,3	2.565	2,1	32.131	26,2	**122.863**
Desconhecido	86.251	25,1	2.230	0,6	88.229	25,6	8.853	2,6	158.671	46,1	**344.234**
Enforcamento	17.295	9,0	-	0,0	2.424	1,3	155.739	81,4	15.912	8,3	**191.370**
Envenenamento	909	1,5	-	0,0	13.792	22,4	32.852	53,3	14.134	22,9	**61.687**
Fogo	5.485	12,8	12	0,0	26.380	61,6	4.491	10,5	6.430	15,0	**42.798**
Impacto	861	0,3	-	0,0	275.160	94,3	9.228	3,2	6.581	2,3	**291.830**
PAF[1]	913.081	90,8	17.174	1,7	8.517	0,8	29.811	3,0	36.927	3,7	**1.005.510**
Perfurante	191.257	93,1	270	0,1	1.427	0,7	4.118	2,0	8.351	4,1	**205.423**
Veículo	1.860	0,2	-	0,0	957.939	99,7	1.035	0,1	347	0,0	**961.181**
Total	**1.306.155**	**38,5**	**19.822**	**0,6**	**1.523.715**	**44,9**	**251.566**	**7,4**	**294.752**	**8,7**	**3.396.010**

Fonte: MS/SVS/Dados/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Nota: [1] Perfuração por arma de fogo.

TABELA A.2
Homicídios registrados e projetados no Brasil (1996-2021)

Ano	Homicídios registrados (A)	Homicídios ocultos (B)	Homicídios projetados (C)	Mortes externas indeterminadas (D)	B/D (%)	C/A ▲(%)
1996	38.929	5.028	43.957	9.820	51,2	12,9
1997	40.531	4.628	45.159	9.159	50,5	11,4
1998	41.974	6.711	48.685	13.118	51,2	16,0
1999	42.947	5.751	48.698	10.769	53,4	13,4
2000	45.433	5.982	51.415	11.934	50,1	13,2
2001	48.032	5.526	53.558	11.520	48,0	11,5
2002	49.816	6.639	56.455	12.557	52,9	13,3
2003	51.534	5.316	56.850	11.101	47,9	10,3
2004	48.909	5.049	53.958	11.597	43,5	10,3
2005	48.136	4.881	53.017	11.269	43,3	10,1
2006	49.704	3.694	53.398	9.147	40,4	7,4
2007	48.219	5.245	53.464	11.367	46,1	10,9
2008	50.659	5.406	56.065	12.056	44,8	10,7
2009	52.043	5.827	57.870	13.253	44,0	11,2
2010	53.016	3.471	56.487	9.703	35,8	6,5
2011	52.807	3.669	56.476	10.353	35,4	6,9
2012	57.045	3.873	60.918	10.051	38,5	6,8
2013	57.396	3.479	60.875	9.788	35,5	6,1
2014	60.474	3.667	64.141	9.468	38,7	6,1
2015	59.080	3.766	62.846	9.810	38,4	6,4
2016	62.517	4.019	66.536	10.274	39,1	6,4
2017	65.602	3.582	69.184	9.799	36,6	5,5
2018	57.956	5.238	63.194	12.310	42,6	9,0
2019	45.503	7.366	52.869	16.648	44,2	16,2
2020	49.868	5.602	55.470	14.038	39,9	11,2
2021	47.847	5.152	52.999	13.843	37,2	10,8

Fonte: MS/SVS/Dados/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.

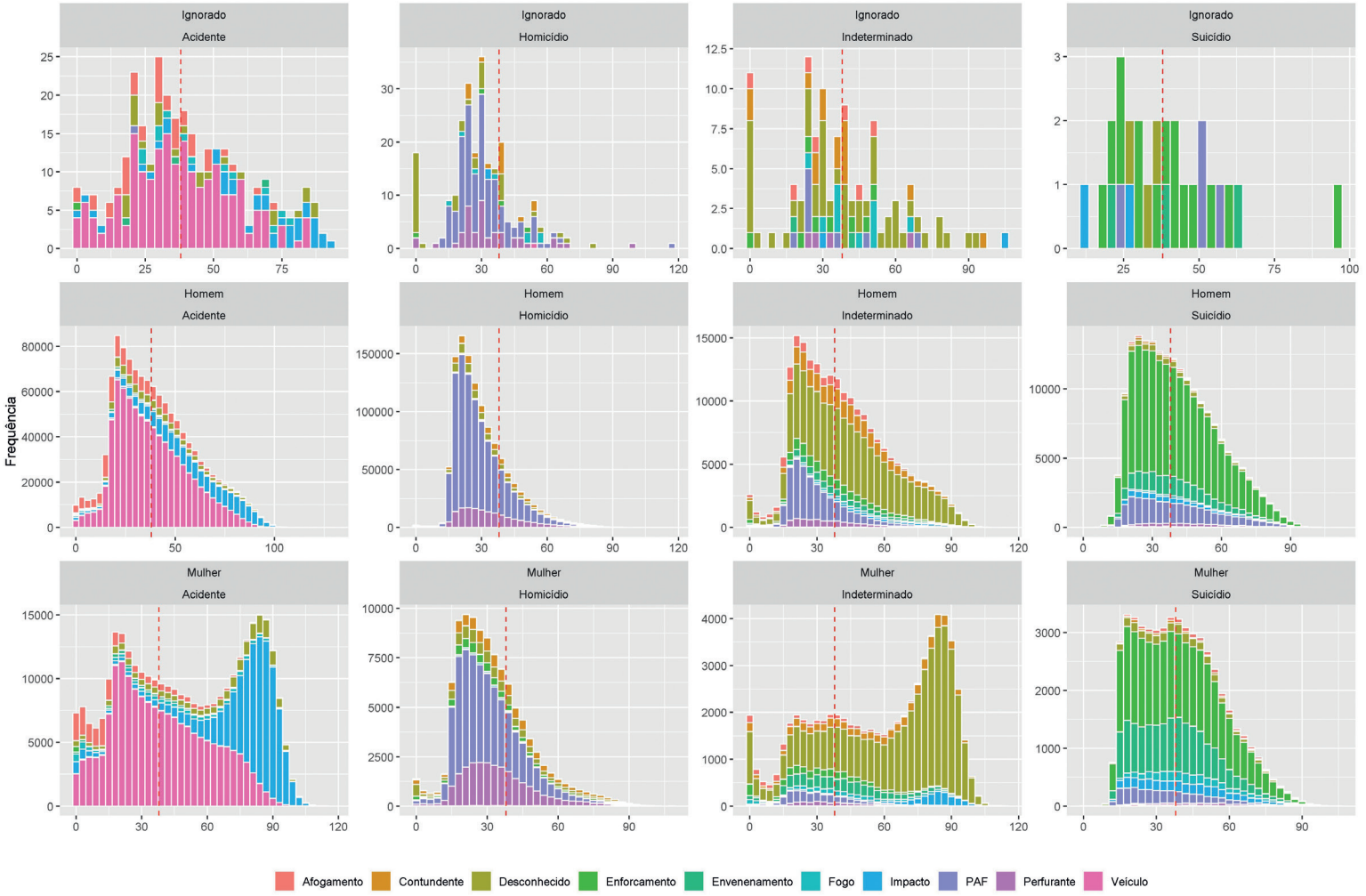
TABELA A.3
Similaridades entre homicídios ocultos e homicídios redistribuídos

Unidades da Federação (UFs)	Erro médio absoluto (MAE)	Raiz do erro quadrático médio (RMSE)	Desvio médio absoluto (MAD)
Acre	3,4	4,5	2,6
Alagoas	5,4	6,8	5,4
Amapá	2,7	4,1	1,9
Amazonas	10,8	12,9	6,5
Bahia	235,1	266,9	176,3
Ceará	148,6	187,2	135,2
Espírito Santo	19,4	25,7	14,0
Distrito Federal	3,5	5,0	2,5
Goiás	47,2	59,8	13,2
Maranhão	31,6	35,8	23,1
Mato Grosso	16,7	20,9	15,5
Mato Grosso do Sul	7,9	9,3	4,5
Minas Gerais	306,0	324,4	147,6
Pará	41,6	46,9	25,4
Paraíba	18,1	22,9	13,6
Paraná	53,6	63,1	46,0
Pernambuco	268,2	294,1	155,8
Piauí	17,4	21,8	14,5
Rio de Janeiro	356,8	399,1	205,5
Rio Grande do Norte	49,3	56,2	38,8
Rio Grande do Sul	73,8	82,5	50,2
Rondônia	15,8	20,3	10,6
Santa Catarina	15,1	19,8	13,1
São Paulo	477,6	619,0	453,8
Sergipe	42,5	49,6	20,4
Tocantins	5,7	8,5	5,2
Roraima	6,2	9,6	5,0

Fonte: MS/SVS/Dados/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.

GRÁFICO A.1

Histogramas de idade da vítima e instrumento do óbito, por sexo e intencionalidade do óbito

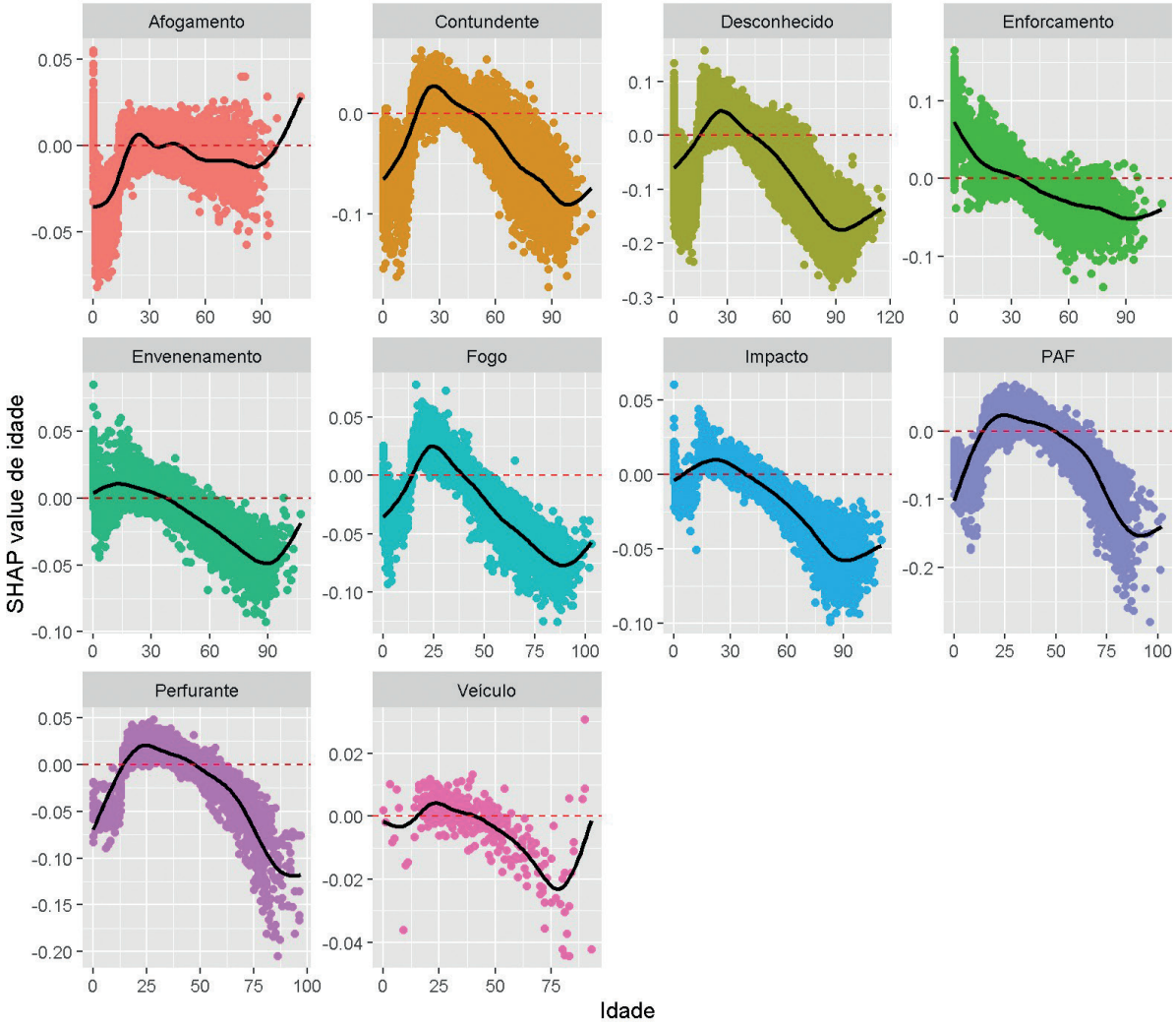


Elaboração dos autores.

Obs.: Ilustração cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

GRÁFICO A.2

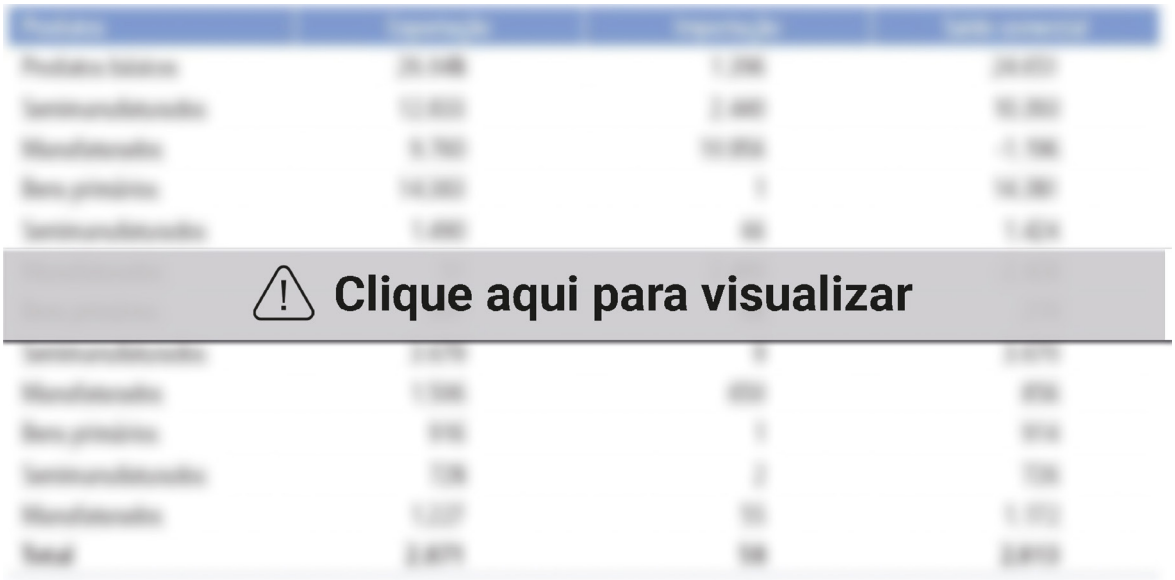
***SHAP value* de idade, por instrumento do óbito**



Elaboração dos autores.

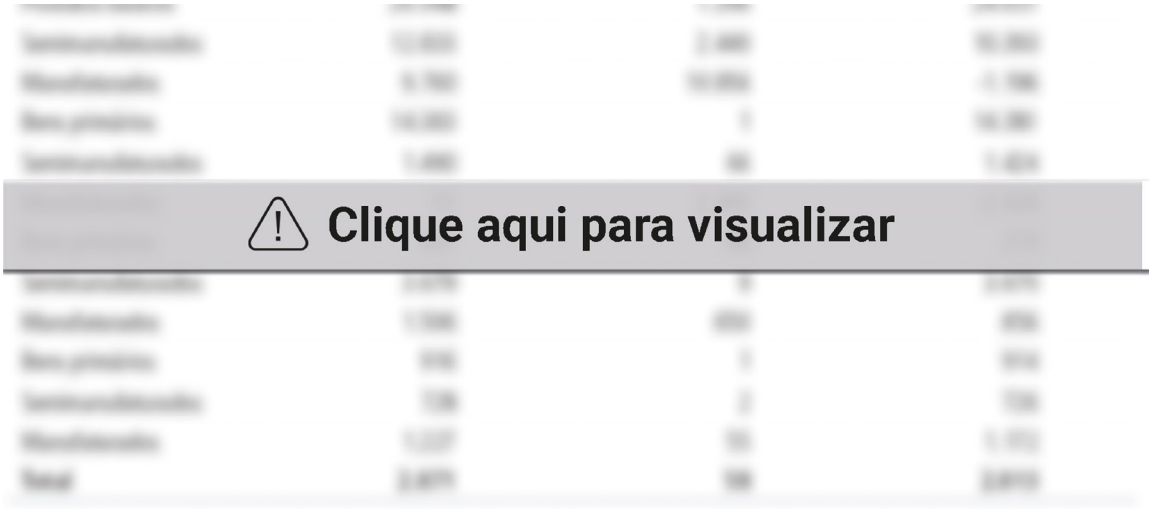
Obs.: Ilustração cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial).

TABELA A.4
Homicídios registrados nas UFs (1996-2021)



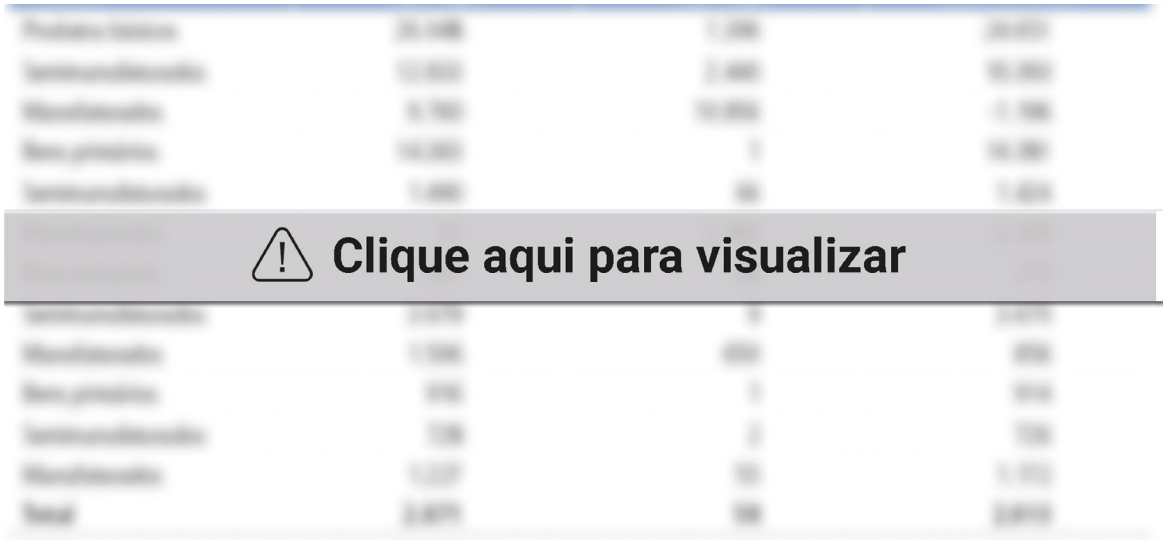
Fonte: MS/SVS/Dados/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Obs.: Para a visualização da tabela em sua integralidade, favor acessar: https://repositorio.ipea.gov.br/bitstream/11058/14016/1/TABELA_A.4.xlsx (nota do Editorial).

TABELA A.5
Homicídios ocultos nas UFs (1996-2021)



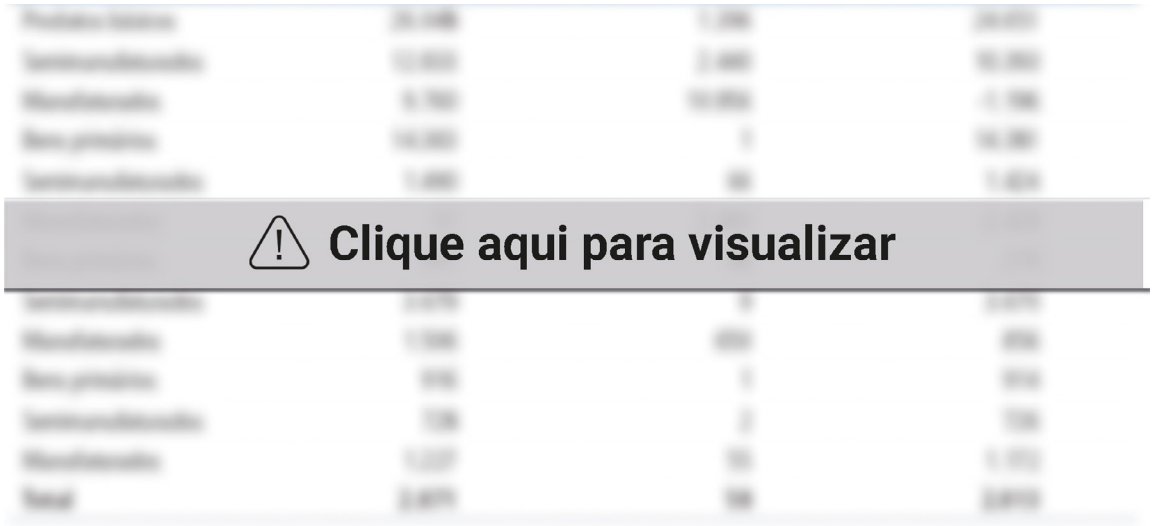
Fonte: MS/SVS/Dados/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Obs.: Para a visualização da tabela em sua integralidade, favor acessar: https://repositorio.ipea.gov.br/bitstream/11058/14016/2/TABELA_A.5.xlsx (nota do Editorial).

TABELA A.6
Homicídios projetados nas UFs (1996-2021)



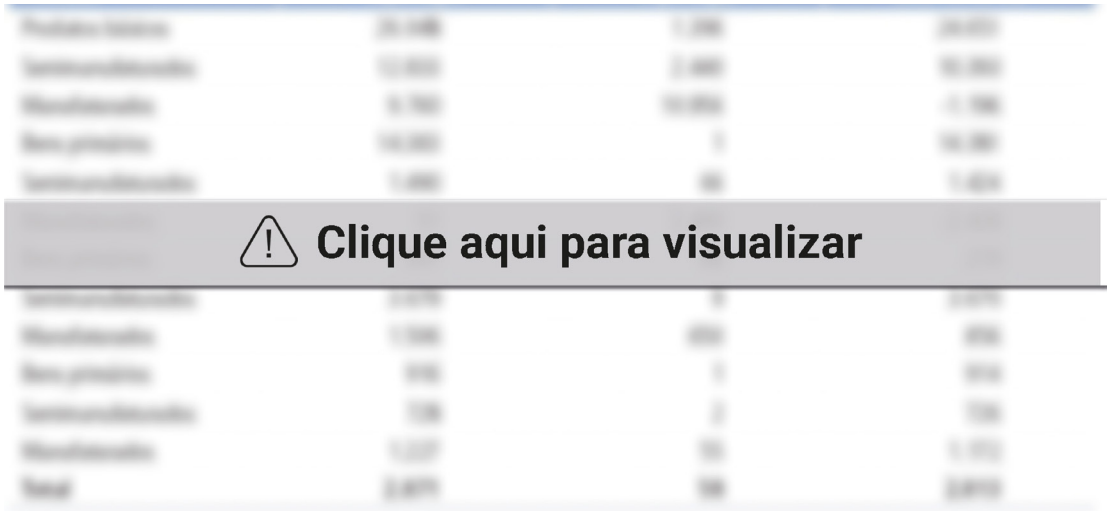
Fonte: MS/SVS/Dados/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Obs.: Para a visualização da tabela em sua integralidade, favor acessar: https://repositorio.ipea.gov.br/bitstream/11058/14016/3/TABELA_A.6.xlsx (nota do Editorial).

TABELA A.7
Taxas de homicídios registrados nas UFs (1996-2021)
(Por 100 mil hab.)



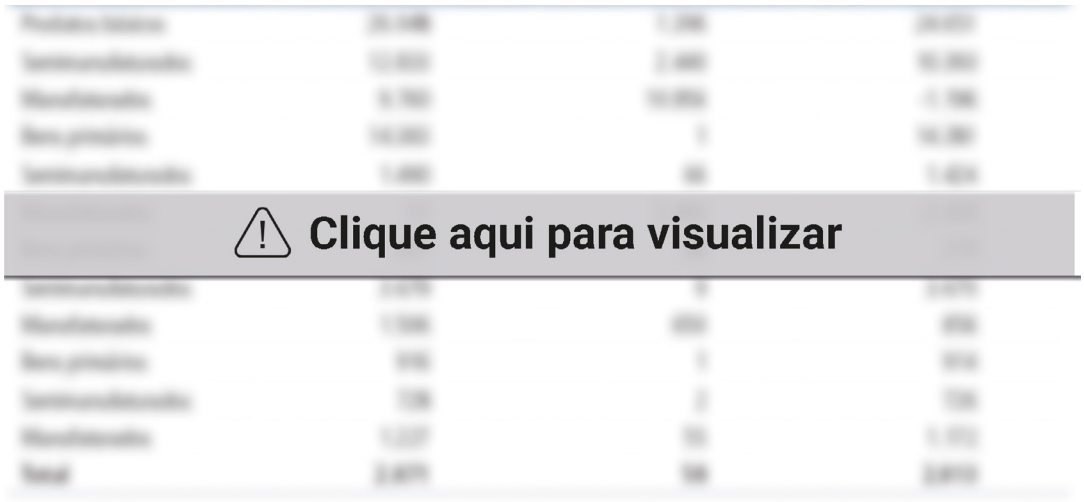
Fonte: MS/SVS/Dados/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Obs.: Para a visualização da tabela em sua integralidade, favor acessar: https://repositorio.ipea.gov.br/bitstream/11058/14016/4/TABELA_A.7.xlsx (nota do Editorial).

TABELA A.8
Taxas de homicídios ocultos nas UFs (1996-2021)
(Por 100 mil hab.)



Fonte: MS/SVS/Dados/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Obs.: Para a visualização da tabela em sua integralidade, favor acessar: https://repositorio.ipea.gov.br/bitstream/11058/14016/5/TABELA_A.8.xlsx (nota do Editorial).

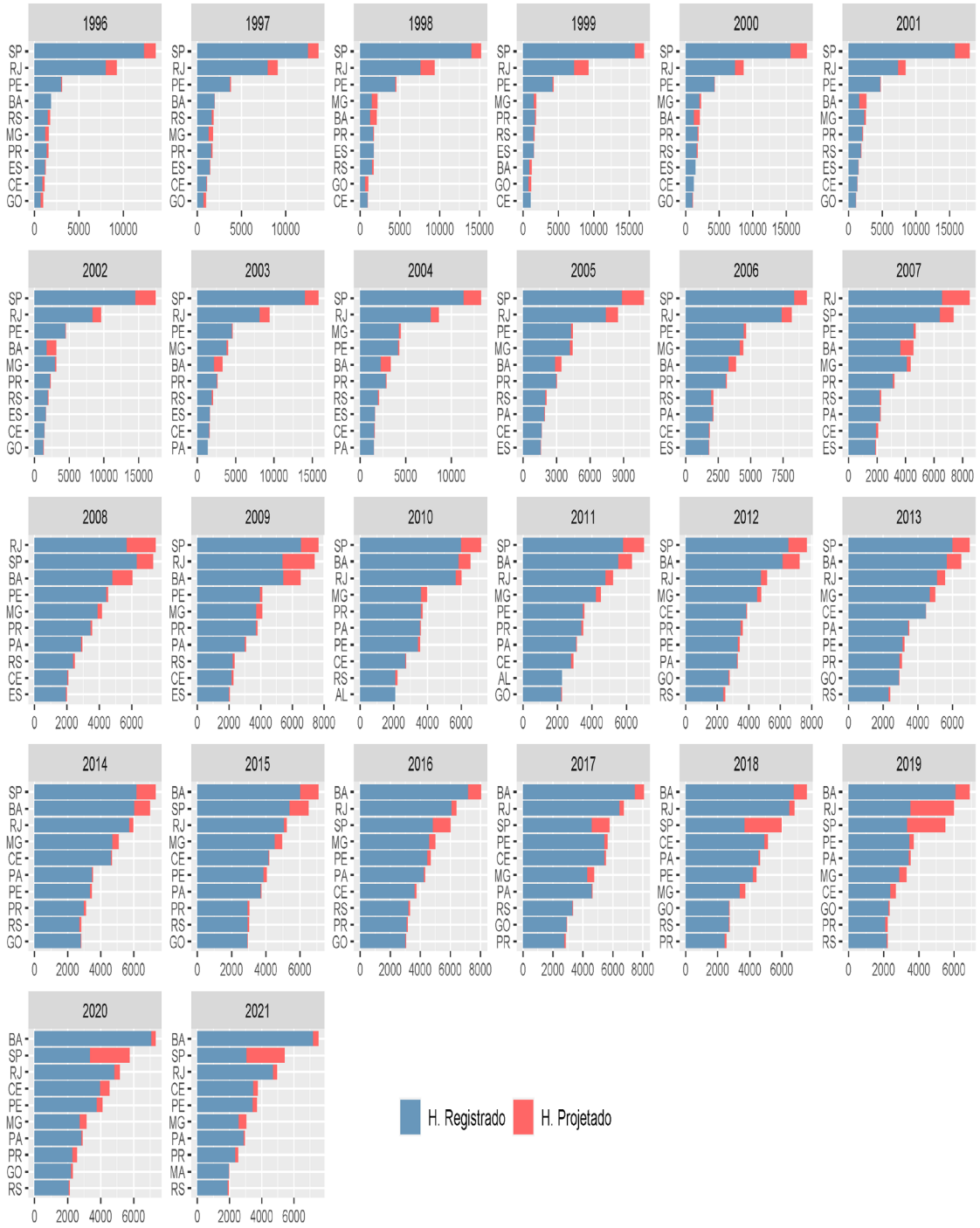
TABELA A.9
Taxas de homicídios nas UFs (1996-2021)
(Por 100 mil hab.)



Fonte: MS/SVS/Dados/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.
Obs.: Para a visualização da tabela em sua integralidade, favor acessar: https://repositorio.ipea.gov.br/bitstream/11058/14016/6/TABELA_A.9.xlsx (nota do Editorial).

GRÁFICO A.3

Valores absolutos de homicídios registrados e projetados, por UF (1996-2021)



Fonte: MS/SVS/Dados/SIM. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/.
Elaboração dos autores.

Obs.: Ilustração cujos leiaute e textos não puderam ser padronizados e revisados em virtude das condições técnicas dos originais (nota do Editorial)

QUADRO A.1
Modelos de classificação

Modelo/especificação	Descrição
Logit A formulação geral do *logit*: $\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \sum_{j=1}^{p}\beta_j x_j$ em que: • *p(x)* é a probabilidade condicional de sucesso (homicídio oculto), dado o vetor de preditores *x*; • β_0 é o intercepto; • β_j são os coeficientes associados às variáveis previsoras x_j; e • *p* é o número de variáveis previsoras.	Suponhamos uma variável dependente binária *Y* (em que *Y* = 1 se homicídio, e *Y* = 0 se acidente/suicídio) e um conjunto de *p* variáveis previsoras. Estamos interessados em modelar a probabilidade condicional Pr(Y = 1 | X = x). Os parâmetros do modelo são estimados utilizando-se máxima verossimilhança.
Logit penalizado (*lasso, ridge* e *elastic-net*) Em Zou e Hastie (2005), os autores acrescentam ao *logit* a combinação de duas penalizações, isto é, L1 (*lasso logit*) e L2 (*ridge logit*), resultando no seguinte modelo: $\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \sum_{j=1}^{p}\beta_j x_j +$ $\lambda\sum_{j=1}^{p}\left(\alpha\lvert\beta_j\rvert + (1-\alpha)\beta_j^2\right)$ • λ é a ponderação que controla as penalizações L1 e L2; e • α é o parâmetro de penalização.	O parâmetro α assume valores entre 0 e 1 e controla o equilíbrio entre as penalizações 1 (norma L1) e L2 (norma euclidiana) no modelo. 1) Quando $\alpha = 1\ e\ \lambda = 1$, o modelo *logit elastic net* se reduz ao modelo *lasso logit*. Nesse caso, apenas a penalização L1 é aplicada. Isso promove a exclusão de variáveis (*feature selection*) com pouco poder preditivo, ao forçar alguns coeficientes β_j a serem exatamente 0. É útil para redução de dimensionalidade e seleção de características. O valor ótimo do hiperparâmetro α é determinado através de 10-*fold cross-validation* na base de treinamento. Os parâmetros do modelo são estimados utilizando-se máxima verossimilhança. 2) Se $\alpha = 0\ e\ \lambda = 1$, o *logit elastic net* torna-se *ridge logit*. A inclusão do termo de penalização *ridge* na função objetivo força os coeficientes de variáveis pouco importantes a serem pequenos, reduzindo assim a variância do modelo e evitando o *overfitting*. Nesse caso, apenas a penalização L2 é aplicada, e não há penalização L1. É útil quando todas as variáveis previsoras são consideradas importantes. O valor ótimo do hiperparâmetro α é determinado através de 10-*fold cross-validation* na base de treinamento. Os parâmetros do modelo são estimados utilizando-se máxima verossimilhança. 3) Quando $0 < \lambda < 1$ e $\alpha \neq 0$, o modelo *elastic net logit* combina as penalizações L1 e L2 para obter um equilíbrio entre seleção de características (*lasso*) e regularização (*ridge*). O valor específico de α determina a proporção relativa das penalizações L1 e L2. Um valor maior de α dará mais peso à penalização L1, enquanto um valor menor de α dará mais peso à penalização L2. Os valores ótimos dos hiperparâmetros $\alpha\ e\ \lambda$ são determinados através de 10-*fold cross-validation* na base de treinamento. Os parâmetros do modelo são estimados utilizando-se máxima verossimilhança.

(Continua)

(Continuação)

Modelo/especificação	Descrição
Decision tree Suponhamos um conjunto de observações (*X*,*Y*), em que: • *X* é uma matriz de dimensões *n x p*, onde *n* é o número de observações e *p* é o número de variáveis previsoras. • *Y* é um vetor binário de dimensão *n* que representa as classes das observações. Nesse caso, *Y* será homicídio (*Y*=1) ou acidente/suicídio (*Y*=0). • O objetivo do modelo de árvore de decisão é aprender uma árvore que divida o espaço de características de *X* em regiões (ou nós), de forma a minimizar o índice de Gini, uma medida de impureza dos nós. O índice de Gini é definido como $Gini(D) = 1 - \sum_{i=1}^{2}(p_i)^2$, em que p_i é a proporção de observações da classe *i* no conjunto de dados (D), isto é, o nó investigado. Quanto menor o índice de Gini, mais puro é o conjunto de dados, o que significa que as amostras pertencem principalmente a uma única classe. A construção da árvore ocorre conforme a seguir descrito. 1) Começamos com um nó raiz que contém todas as observações. 2) O algoritmo calcula o índice de Gini nesse nó e, em seguida, procura por todas as divisões possíveis nas variáveis previsoras, escolhendo a divisão que minimiza o índice de Gini para os nós filhos. A árvore de decisão escolhe a variável previsora *P* e o valor de corte *V* que minimizam a impureza ponderada dos grupos resultantes, conforme calculado pelo índice de Gini. O critério de divisão é: Gini_split (P,V) = $(\mathrm{P},\mathrm{V}) = \frac{n_1}{n} * Gini(D_1) + \frac{n_2}{n} * Gini(D_2)$, em que n_1 e n_2 são o número de amostras nos grupos resultantes D_1 e D_2. 3) Repetimos o passo 2 recursivamente para cada nó filho, até se atingir um critério de parada, como profundidade máxima da árvore (*tree_depth*) ou número mínimo de observações em um nó (*min_n*). 4) Atribuímos uma classe às folhas da árvore, com base na classe majoritária dessa folha.	A partir das variáveis previsoras, a árvore de decisão compartimenta o espaço de previsores em subgrupos homogêneos, maximizando a pureza, isto é, a proporção de uma classe (homicídio ou acidente/suicídio), em cada subconjunto, até atingir subgrupos com no mínimo *min_n* observações ou atingir profundidade máxima da árvore (*tree_depth*). A divisão é realizada recursivamente e, neste trabalho, a árvore particiona os dados, ao minimizar o índice Gini, sugerido em (Breiman *et al*., 1984). De modo a evitar elaboração de árvore sujeita a *overfitting*, os subgrupos menos informativos são excluídos, isto é, utilizamos o hiperparâmetro (*cost_complexity*), em que a pureza dos subgrupos é penalizada de acordo com o número total de subgrupos finais. *Profundidade máxima da árvore* (*tree_depth*): Esse hiperparâmetro controla a profundidade máxima da árvore. Ele define o número máximo de camadas da árvore a partir do nó raiz. Por exemplo, definindo-se *tree_depth* como 3, a árvore terá no máximo três níveis de profundidade a partir do nó raiz. Controlar a profundidade da árvore é uma maneira de evitar *overfitting*, pois árvores profundas podem se ajustar demais aos dados de treinamento. *Número mínimo de observações por nó* (*min_n*): Esse hiperparâmetro define o número mínimo de observações que um nó deve ter para que uma divisão adicional seja tentada. Se o número de observações em um nó for menor do que esse valor mínimo, a divisão não ocorrerá, o que pode ajudar a evitar subdivisões excessivas em regiões com poucos dados. *Custo de complexidade* (*cost_complexity*) O *cost_complexity* é usado para calcular o custo associado a cada nó da árvore. São excluídos da árvore de decisão nós cujo custo excede o limiar definido em *cost_complexity*. Aumentar o valor de *cost_complexity* leva a árvores menos complexas, pois os nós com *cost_complexity* maiores são excluídos primeiro. A combinação ótima de hiperparâmetros *tree_depth* e *min_n* foi selecionada através de 10-*fold cross-validation* na base de treinamento.
Bagging (*bootstrap aggregating*) *com árvores de decisão* Suponhamos um conjunto de observações (*X*,*Y*), em que: • *X* é uma matriz de dimensões *n x p*, em que *n* é o número de observações e *p* é o número de variáveis previsoras. • *Y* é um vetor binário de dimensão *n* que representa as classes das observações. Nesse caso, *Y* será homicídio (*Y* = 1) ou acidente/suicídio (*Y* = 0). A técnica de *bagging* com árvores de decisão funciona da seguinte forma: • *bootstrap*: gere cem conjuntos de treinamento de tamanho *n* por meio de amostragem com reposição; • *construção de árvores de decisão*: para cada conjunto amostrado, construa uma árvore de decisão usando o índice de Gini como critério, ao particionar os dados. A profundidade da árvore é controlada pelos hiperparâmetros profundidade máxima (*tree_depth*) ou número mínimo de observações por nó (*min_n*); e • *classificação por maioria*: classifique a observação de interesse, considerando todas as árvores construídas. Em problemas de classificação, as árvores votam em uma classe e a classe mais frequente entre as árvores é escolhida como a classe prevista.	O *bagging* faz parte dos métodos de assembleia, isto é, abordagem que combina a previsão de diversos modelos preditivos – nesse caso árvores de decisão –, ao tentar elaborar um modelo preditivo mais poderoso. A árvore de decisão está sujeita a diversos problemas, tais como elevada variância, sensibilidade aos dados de treinamento e *overffiting*. A combinação de previsões executadas no *bagging* reduz a variância e aumenta a *performance* das previsões. A ideia é que, ao se criarem várias árvores de decisão e serem combinadas suas previsões, os erros individuais de cada árvore são reduzidos, aumentando a capacidade de generalização do modelo. Assim, dado um conjunto de observações D, o *bagging*, nesse caso, gera cem conjuntos de observações amostrados aleatoriamente $\{D_1, D_2, \ldots, D_B\}$, com reposição. Para cada conjunto D_b, elabora-se uma árvore de decisão que particiona os dados, minimizando-se o índice de Gini. Em seguida, a observação de interesse é classificada de acordo com a classe mais frequente nas cem árvores de decisão elaboradas (*majority-vote rule*). A combinação ótima de hiperparâmetros *tree_depth* e *min_n* foi selecionada através de 10-*fold cross-validation* na base de treinamento.

(Continua)

(Continua)

(Continuação)

Modelo/especificação	Descrição
Random forest Seguindo Breiman (2001) e Hastie, Tibshirani e Friedman (2009), o *random forest* segue algoritmo abaixo. 1) Para *b* = 1 até B conjuntos de treinamento: a) desenhe uma amostra *bootstrap* $\mathbf{Z}^*$ de tamanho *N* a partir dos dados de treinamento. b) cresça um *random forest* T_b na amostra *bootstrap*, repetindo recursivamente os seguintes passos para cada nó terminal da árvore, até que o tamanho mínimo de nó *n_min* seja alcançado: • selecione aleatoriamente *mtry* variáveis das *p* variáveis; • escolha o melhor ponto de divisão/variável entre as *mtry* selecionadas; e • divida o nó em dois nós filhos. 2) Produza a assembleia de árvores $\{T_b\}_1^B$. Para realizar classificação da nova observação x: Seja $\hat{C}_b(x)$ a classe prevista na b-ésima *andom forest*. Então, $\hat{C}_{rf}^B(x) = voto\ majoritário\ \{\hat{C}_b(x)\}_1^B$.	O *random forest* é uma assembleia de previsores utilizando *decision trees* aplicadas a um aleatório subconjunto dos dados (Biau, 2012). O *random forest* busca reduzir a correlação entre as árvores de decisão existentes no *bagging decision tree* (Breiman, 2001). Nessa abordagem, variáveis com elevado poder preditivo serão repetidamente utilizadas no *split* inicial das árvores de decisão, tornando as árvores correlacionadas, isto é, semelhantes entre si. A elevada correlação entre as árvores indica que elas estão aprendendo os mesmos padrões de dados. Isso significa que, se uma árvore comete um erro em uma determinada região do espaço das variáveis, é provável que outras árvores cometam o mesmo erro nessa região. Nesse caso, o *bagging* não aumenta a capacidade de generalização do modelo. O *random forest* soluciona esse problema, ao forçar cada divisão a considerar apenas um subconjunto aleatório de previsores. Similar ao *bagging*, o *random forest* utilizado neste trabalho constrói cem árvores de decisão a partir de conjuntos de observações amostradas aleatoriamente $\{D_1\ D_2,\ldots,\ D_B\}$ com reposição. No entanto, durante a construção de cada árvore de decisão, somente um subconjunto *mtry*, aleatoriamente selecionado, de variáveis previsoras é considerado, ao determinar a divisão em cada nó da árvore. Utilizando-se somente as *mtry* selecionadas, é feita a divisão naquele nó. Isso força as árvores de decisão a se basearem em diferentes conjuntos de características, tornando-as menos correlacionadas. A combinação de árvores independentes, cada uma treinada em uma amostra diferente e com variáveis preditivas diferentes, ajuda a reduzir a correlação e a melhorar a capacidade de generalização da assembleia de árvores de decisão. Portanto, o *random forest* constrói cem árvores de decisão não correlacionadas (*de-correlated trees*) e, então, quando uma nova observação precisa ser classificada, cada árvore na floresta faz uma previsão com base em suas características. A classe final da nova observação é determinada por votação majoritária. A combinação ótima de hiperparâmetros *mtry* e *min_n* foi selecionada através de 10-*fold cross-validation* na base de treinamento.

Elaboração dos autores.

REFERÊNCIAS

BIAU, G. Analysis of a random forests model. **Journal of Machine Learning Research**, v. 13, p. 1063-1095, 2012.

BREIMAN, L. *et al*. **Classification and regression trees**. Londres: Routledge, 1984.

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

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. Random forests. *In*: HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning**. 2. ed. Nova York: Springer, 2009. p. 587-604.

ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. **Journal of the Royal Statistical Society Series B**, v. 67, n. 2, p. 301-320, abr. 2005.

APÊNDICE B

MÉTRICAS DE CLASSIFICAÇÃO

Como métricas de avaliação, utilizamos o proposto em Fernández *et al*. (2018) e Kuhn e Johnson (2013). O desempenho preditivo, em problemas de classificação binária, pode ser avaliado utilizando-se diversas métricas. Assim, considere-se a matriz de confusão do quadro B.1.

QUADRO B.1
Matriz de confusão

Classe prevista	Classe observada	
	Homicídios	Acidentes/suicídios
Homicídios	TP	FP
Acidentes/suicídios	FN	TN

Elaboração dos autores.

Na matriz de confusão, a diagonal principal representa observações corretamente classificadas, tal que *TP* são verdadeiros positivos, *TN* são verdadeiros negativos, *FP* são falsos positivos e *FN* são falsos negativos. A seguir, são mostradas as métricas de desempenho.

1 ACURÁCIA

$$\text{Acc} = \frac{TP+TN}{N}$$

Na equação, $N = TP + FP + FN + TN$.

A acurácia informa a proporção de observações corretamente classificadas. Apesar da fácil interpretação, a acurácia não impõe custos distintos aos diferentes erros de classificação, isto é, o falso negativo ou falso positivo. Assim, por exemplo, utilizar a acurácia em análises onde o falso negativo impõe custo maior que o falso positivo pode levar a conclusões inadequadas. Outra deficiência da acurácia é não considerar o desbalanceamento de classes. Na presença de desbalanceamento de classes, maximizar a acurácia pode levar a avaliações inadequadas (Kuhn e Johnson, 2013).

2 SENSIBILIDADE/REVOCAÇÃO/TAXA DE ACERTO DA CLASSE POSITIVA/*RECALL*

Sensibilidade = $\frac{TP}{TP+FN}$

Em problemas de classificação, a ocorrência de classe específica pode ser do interesse do pesquisador. A classe de interesse é chamada de classe positiva. Nossa análise considera o homicídio como classe positiva. A sensibilidade informa a proporção de acertos da classe positiva, isto é, a proporção de observações registradas como positivas que são corretamente previstas como positivas (Fernández *et al*., 2018). Elevada sensibilidade indica reduzidos falsos negativos. Os valores estão entre 0, indicando nenhuma sensibilidade, e 1, indicando sensibilidade máxima.

3 PRECISÃO

Precisão = $\frac{TP}{TP+FP}$

Similarmente à sensibilidade, a precisão informa o desempenho preditivo na classe de interesse. Assim, a precisão do modelo informa a proporção de previstos positivos que são observados como positivos. Isto é, a fração de observações corretamente classificadas entre aquelas classificadas como positivas (Fernández *et al*., 2018) ou o erro dos falsos positivos. Os valores estão entre 0, indicando nenhuma precisão, e 1, indicando máxima precisão.

4 F-*MEASURE*

A F-*measure* combina o *trade-off* entre sensibilidade e precisão através de média harmônica ponderada, isto é:

$$F_\beta = (1 + \beta^2)\frac{Precisão * sensibilidade}{(\beta^2 * precisão) + sensibilidade}$$

Na equação, *B* representa a importância relativa entre precisão e sensibilidade. Assim, quanto maior a precisão e o *recall*, mais a F-*measure*. Atribuindo pesos iguais às duas métricas, isto é, $\beta = 1$, teremos a $F_1 measure$:

$$F_1 = 2\frac{Precisão * sensibilidade}{Precisão + sensibilidade}$$

Um modelo acertando todas as previsões tem F_1 igual a 1.

5 COEFICIENTE DE CORRELAÇÃO DE MATTHEWS

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

Ao utilizar as métricas *precisão* e *sensibilidade*, a F-*measure*, interessada na classe positiva, não informa acerca dos falsos negativos e, portanto, não captura toda informação da matriz de confusão. Uma métrica capaz de utilizar toda informação da matriz de confusão é o coeficiente de correlação de Matthews (MCC).

O MCC elabora o coeficiente de correlação entre duas variáveis binárias, nesse caso, a variável da classe prevista e a variável da classe observada. Quanto maior a correlação, melhor a previsão. Entre as métricas utilizadas, é a única que considera as quatro dimensões da matriz de confusão, e está limitada entre -1 e 1. Isto é, -1 indica que todas as previsões estão erradas, 0 indica correlação aleatória e 1 indica que todas as previsões estão certas.

6 ESPECIFICIDADE/TAXA DE ACERTO DA CLASSE NEGATIVA

Especificidade = $\frac{TN}{TN+FP}$

A especificidade indica a proporção de observações negativas corretamente previstas como negativas. Nossa análise considera os acidentes/suicídios como classe negativa. Assim, indica a proporção de acidentes/suicídios incorretamente previstos como homicídios. Elevada especificidade indica reduzido número de falsos positivos.

7 YOUDEN'S J INDEX

$$J\ index = Sensibilidade + Especificidade - 1$$

O índice *J* de Youden mensura a proporção de observações corretamente prevista nas classes positiva e negativa. Com valores entre 0 e 1, a métrica serve como sumário da magnitude de erros nas duas classes. Quando o valor é 1, não ocorre falso positivo nem falso negativo, e as previsões são perfeitas.

8 BALANCED ACCURACY

BAcc = $\frac{Sensibilidade+Especificidade}{2}$

Balanced accuracy é a média entre sensibilidade e especificidade. Corrige problemas da acurácia em bases desbalanceadas.

9 *RECEIVING OPERATING CHARACTERISTIC* E *AREA UNDER THE RECEIVER OPERATOR CURVE*

As métricas anteriores utilizam somente informação da classe prevista (homicídio ou acidente/suicídio), não considerando a probabilidade estimada em cada observação e o *threshold* de classificação. A curva *receiving operating characteristic* (ROC) apresenta variações no desempenho preditivo do modelo ao conectar, no plano cartesiano, duas métricas, isto é, 1-especificidade (taxa de falso positivo) no eixo x e sensibilidade (taxa de verdadeiro positivo) no eixo y, em diferentes *thresholds*. Ao considerar diferentes valores do *threshold*, investigamos o *trade-off* entre a 1-especificidade (taxa de falso positivo) e a sensibilidade (taxa de verdadeiro positivo), elaborando novas estimações em cada *threshold* investigado. *Grosso modo*, a curva ROC é a união dessas estimações no gráfico cartesiano. Por ser uma função de sensibilidade e especificidade, a área sob a curva ROC (*area under the ROC curve* – AUC-ROC) é indiferente a desbalanceamento de classes (Fawcett, 2006).

Métrica capaz de sintetizar o desempenho em diferentes *thresholds*, a AUC produz valores entre 0 e 1, e pode ser interpretada como a probabilidade de o modelo de classificação classificar um caso positivo, aleatoriamente escolhido, como tendo uma probabilidade de ser positivo mais alta do que um caso negativo (Fawcett, 2006). Em geral, o modelo com a maior área abaixo da curva ROC é considerado aquele de melhor desempenho preditivo.

REFERÊNCIAS

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861-874, jun. 2006.

FERNÁNDEZ, A. *et al*. Performance measures. *In*: FERNÁNDEZ, A. *et al*. **Learning from imbalanced data sets**. Cham: Springer, 2018. p. 47-61.

KUHN, M.; JOHNSON, K. Measuring performance in classification models. *In*: KUHN, M.; JOHNSON, K. **Applied predictive modeling**. Nova York: Springer, 2013. p. 247-273.

Ipea – Instituto de Pesquisa Econômica Aplicada

EDITORIAL

Coordenação
Aeromilson Trajano de Mesquita

Assistentes da Coordenação
Rafael Augusto Ferreira Cardoso
Samuel Elias de Souza

Supervisão
Ana Clara Escórcio Xavier
Everson da Silva Moura

Revisão
Alice Souza Lopes
Amanda Ramos Marques Honorio
Barbara de Castro
Brena Rolim Peixoto da Silva
Cayo César Freire Feliciano
Cláudio Passos de Oliveira
Nayane Santos Rodrigues
Clícia Silveira Rodrigues
Olavo Mesquita de Carvalho
Reginaldo da Silva Domingos
Jennyfer Alves de Carvalho (estagiária)
Katarinne Fabrizzi Maciel do Couto (estagiária)

Editoração
Anderson Silva Reis
Augusto Lopes dos Santos Borges
Cristiano Ferreira de Araújo
Daniel Alves Tavares
Danielle de Oliveira Ayres
Leonardo Hideki Higa
Natália de Oliveira Ayres

Capa
Aline Cristine Torres da Silva Martins

Projeto Gráfico
Aline Cristine Torres da Silva Martins

The manuscripts in languages other than Portuguese published herein have not been proofread.

Ipea – Brasília
Setor de Edifícios Públicos Sul 702/902, Bloco C
Centro Empresarial Brasília 50, Torre B
CEP: 70390-025, Asa Sul, Brasília-DF

Missão do Ipea

Aprimorar as políticas públicas essenciais ao desenvolvimento brasileiro por meio da produção e disseminação de conhecimentos e da assessoria ao Estado nas suas decisões estratégicas.

ipea Instituto de Pesquisa Econômica Aplicada

MINISTÉRIO DO PLANEJAMENTO E ORÇAMENTO

GOVERNO FEDERAL
BRASIL
UNIÃO E RECONSTRUÇÃO