

Mapa dos homicídios ocultos no Brasil entre 1996 e 2021

The map of hidden homicides in Brazil between 1996 and 2021

Mapa de los homicidios ocultos en Brasil entre 1996 y 2021

Daniel Ricardo de Castro Cerqueira <https://orcid.org/0000-0002-4083-9535>¹; Gabriel de Oliveira Accioly Lins <https://orcid.org/0000-0002-9360-2902>²

Resumo Interessados em mensurar a criminalidade violenta utilizam o número de homicídios, a partir de hipótese, nunca verificada, de que a gravidade do crime implica subnotificação desprezível. No Brasil, o Sistema de Informações sobre Mortalidade (SIM) é fonte habitualmente utilizada na aferição do número de homicídio. Entre 1996 e 2021, dentre as mortes por causa externa, 8,2% são classificadas como mortes por causa externa de intenção indeterminada (MVCI), indicando possibilidade de subnotificação de homicídios. A partir dos microdados sobre morte por causa externa registradas no SIM, algoritmo de aprendizado supervisionado aprende as características de homicídios, acidentes e suicídios e estima probabilidade de as MVCI serem na realidade homicídio incorretamente registrados como MVCI, permitindo-nos estimar o número de homicídios ocultos no SIM. O resultado encontrado sugere que, entre 1996 e 2021, não foram contabilizados 128.567 homicídios ocultos ou 43,6% das MVCI são na realidade homicídios. Evidência contrária à hipótese de reduzida subnotificação, implicando impactos significativos em avaliações de políticas públicas, procedimentos econométricos e necessidade de contabilização destes homicídios ocultos em medições da criminalidade.

Palavras-chave Aprendizado de Maquinas, Homicídio, Crime, Subnotificação, DATASUS

Abstract Those interested in measuring violent crime use the number of homicides, based on the never-verified assumption that the seriousness of the crime implies negligible underreporting. In Brazil, the Mortality Information System (SIM) is the source usually used to measure the number of homicides. Between 1996 and 2021, among deaths from external causes, 8.2% were classified as deaths from external causes of undetermined intent (MVCI), indicating the possibility of homicide underreporting. From the microdata on external cause deaths recorded in the SIM, a supervised learning algorithm learns the characteristics of homicides, accidents and suicides and estimates the probability that MVCI are homicides incorrectly recorded as MVCI, allowing us to estimate the number of shadow homicides in the SIM. The result suggests that between 1996 and 2021, 128,567 shadow homicides were not accounted for, or 43.6% of MVCI are actually homicides. Evidence contrary to the hypothesis of reduced underreporting, implying significant impacts on public policy evaluations, econometric procedures and the need to account for these hidden homicides in crime measurements.

Key words Machine Learning, Homicide, Crime, Underreporting, DATASUS

Resumen Quienes se interesan en medir los delitos violentos utilizan el número de homicidios, basándose en la hipótesis, nunca verificada, de que la gravedad del delito implica un subregistro insignificante. En Brasil, el Sistema de Información de Mortalidad (SIM) es la fuente comúnmente utilizada para medir el número de homicidios. Entre 1996 y 2021, el 8,2% de las muertes por causas externas se clasificaron como muertes por causas externas de intención indeterminada (MVCI), lo que indica la posibilidad de subregistro. Utilizando microdatos sobre muertes por causas externas registrados en el SIM, un algoritmo de aprendizaje supervisado aprende las características de homicidios, accidentes y suicidios y estima la probabilidad de que las MVCI sean en realidad homicidios registrados incorrectamente como MVCI, lo que permite estimar el número de homicidios ocultos en el SIM. Los resultados sugieren que, entre 1996 y 2021, 128.567 homicidios ocultos no se reportaron, o que el 43,6% de las MVCI fueron en realidad homicidios. Esta evidencia contradice la hipótesis de reducción del subregistro, lo que implica impactos significativos en las evaluaciones de políticas públicas, los procedimientos econométricos y la necesidad de contabilizar estos homicidios ocultos en las mediciones de la delincuencia.

Palabras clave Aprendizaje automático, Homicidio, Delincuencia, Subdeclaración, DATASUS

¹ Diretoria de Estudos e Políticas do Estado, das Instituições e da Democracia, Instituto de Pesquisa Econômica Aplicada. Av. Presidente Vargas, 730, 17º andar, Torres 3 e 4, Ed. Bacen, Centro. 20071-001 Rio de Janeiro RJ Brasil. daniel.cerqueira@ipea.gov.br
² Subprograma de Pesquisa para o Desenvolvimento Nacional, Diretoria de Estudos e Políticas do Estado, das Instituições e da Democracia, Instituto de Pesquisa Econômica Aplicada. Rio de Janeiro RJ Brasil.

Introdução

A utilização da taxa de homicídios por 100 mil habitantes como *proxy* da dinâmica criminal aparece justificada e recomendada na literatura na hipótese de pretensa subnotificação reduzida¹. No Brasil, o Sistema de Informação sobre Mortalidade (SIM) do Ministério da Saúde, base de dados em diversos estudos de criminologia (e.g. Atlas da violência²), utiliza a Classificação Internacional de Doenças (CID-10) ao classificar as causas de óbitos. O capítulo XX da CID-10 abarca as mortes por causa externa, isto é, acidente, suicídio, homicídio, sequela médica e morte por causa externa de intenção indeterminada (MVCI). Óbitos classificados com MCVI são eventos onde, a princípio, não é possível distinguir entre homicídio, acidente e suicídio³. Este trabalho investiga a possibilidade de as MVCI serem, na realidade, homicídios não identificados e, portanto, ocultos aos pesquisadores. Possibilidade não considerada em análises sobre criminalidade e com possíveis impactos relevantes em avaliações de políticas públicas.

Entre 1996 e 2021, 8,2% das mortes por causa externa são classificadas como mortes por causa externa de intenção indeterminada (MVCI), isto é, em média o sistema de saúde não consegue identificar a intencionalidade de 11.336 óbitos ao ano, valor superior à média anual de homicídios em São Paulo. Em determinadas unidades da Federação (UFs) o número de MVCI ultrapassou o número de homicídios, impedindo compreensão do real nível de violência. Por exemplo, em São Paulo, entre 2018 e 2021, o número de MVCI foi em média 27,9% superior ao número de homicídios. A média nacional de 6,05 MVCI por 100 mil habitantes aproxima o Brasil ao grupo de países europeus com maior incerteza sobre intencionalidade das mortes por causa externa. Portanto, existe incerteza não desprezível acerca da intencionalidade das mortes por causa externa no Brasil.

Analisando o registro de óbitos em conjunto de países do leste europeu as evidências apresentadas no caso russo⁴ e países bálticos⁵ indicam que parcela significativa das MVCI são na realidade homicídios não registrados. No Brasil as inconsistências em registros do SIM e propostas de retificação são conhecidas⁶. Existem evidências da subenumeração de óbitos no SIM, indicando, de maneira geral, que há óbitos erroneamente classificados, incluindo casos de homicídios que foram registrados como MVCI. As propostas de utilizar técnicas de trabalho de campo e, portanto, necessidade de enfoque lo-

cal, inviabilizam generalização destas propostas aos registros nacionais. Alternativa aos registros oficiais, as *survey* de vitimização pretendem contornar a subnotificação através de questionamentos a população e são desenvolvidas buscando ultrapassar limitações dos dados policiais. No entanto, além da sujeição a viés de preferência revelada⁷, no caso brasileiro, existem somente três pesquisas de abrangência nacional de vitimização⁸, ainda assim, incomparáveis no tempo e focadas em crimes contra propriedade.

Daniel Cerqueira⁹ desenvolveu método capaz de atribuir probabilidade de cada MVCI ser, na realidade, um homicídio. O método identifica os padrões diferenciados de mortes, segundo as características situacionais do incidente e características individuais das vítimas. Neste trabalho aprimoramos a metodologia adotada e ampliamos o período analisado. Assim, a partir dos microdados sobre morte por causa externa registradas no SIM, algoritmo de aprendizado supervisionado aprende as características de homicídios, acidentes e suicídios e estima probabilidade de as MVCI, registradas entre 1996 e 2021, serem na realidade homicídio incorretamente registrados como MVCI, permitindo-nos estimar o número de homicídios ocultos no SIM.

O resultado encontrado sugere que, entre 1996 e 2021, não foram contabilizados 128.567 homicídios ocultos ou 43,6% das MVCI são na realidade homicídios ocultos. O aumento de 9,6% no número total de homicídios, ao considerarmos àqueles ocultos, permite retrato mais acurado da criminalidade violenta, por exemplo, ao indicar fictícia redução de homicídios em algumas UFs e sugere, ao menos no caso brasileiro, violações das hipóteses comumente adotadas no tratamento de variáveis com erro de mensuração, sinalizando a relevância dos resultados e indicando impacto não desprezível em diagnósticos e análises de política pública, incompatível com a esperada subnotificação reduzida preconizada na literatura.

Após esta introdução, a seção dois apresenta a base utilizada e síntese da literatura sobre qualidade dos registros do SIM. A seção três discorre sobre MVCI e apresenta metodologia utilizada. Os homicídios ocultos e projetados são discutidos na seção quarto. Por fim, a última seção resume as conclusões do trabalho.

Metodologia

Por força de lei, nenhum sepultamento deveria ocorrer na ausência de certidão de óbito¹⁰.

No caso de óbitos não naturais, a regulamentações^{11,12} 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. A partir deste exame pericial e em informações prestadas por familiares, indivíduos que socorreram a vítima ou pela polícia, o médico-legista tenta estabelecer a causa básica da morte. Isto é, estabelecer se a morte por causa externa foi ocasionada por: i) acidentes (e.g.; acidente de carro); ii) lesões autoprovocadas intencionalmente (i.e.; suicídio); iii) agressões (e.g.; agressão por meio de arma de fogo); iv) Intervenções legais e operações de guerra; v) mortes cuja intenção é indeterminada (MVCI); vi) Complicações de assistência médica, sequelas médicas ou acidente natural. Com base nas informações apuradas pelo médico-legista, os codificadores das secretarias municipais 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 com morte por causa externa cuja intenção é indeterminada (MVCI).

Assim, considerando óbitos pertencentes ao capítulo XX da CID-10 (i.e. 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 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) morte por causa externa de intenção indeterminada (MVCI) são classificadas no grupo indeterminadas; 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 ulterior, dado impossibilidade deste tipo de óbito ser representado por MVCI.

A partir do código CID-10 registrado na causa básica do óbito a variável instrumento foi criada e informa o instrumento responsável por desencadear o processo mórbido, de modo a capturar dinâmicas associadas às mortes. Com isso, estabeleceram-se dez categorias de instrumento (i.e., envenenamento, enforcamento, afogamento, perfuração por arma de fogo (PAF), instrumento impactante, fogo, instrumento per-

furante, 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 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. Fogo inclui 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 exhibe os grupos de óbitos classificados conforme 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.

A relevância dos óbitos categorizados conforme a CID-10 em pesquisas sobre mortalidade promoveu análises que buscam avaliar a qualidade destes registros e apresentar propostas de sobrepujar as limitações existentes. Em conjunto de ex-repúblicas soviéticas⁵, sugere utilização das MVCI em tentativa de ocultar o número de suicídios e homicídios. Na Rússia⁴, entre 2000 e 2011, 33% das MVCI são, na realidade, homicídios incorretamente registrados. Enquanto na Finlândia¹³ 10% das MVCI são suicídios. Na Argentina¹⁴, entre 1997 e 2018, 28,5% das MVCI são identificadas como homicídios ocultos.

No caso brasileiro, entre 2000 e 2009, o município de Viçosa¹⁵, ao utilizar busca ativa, identificou que 57,7% dos 391 óbitos por causa externa indeterminada são na realidade homicídios. No estado do Rio de Janeiro¹⁶, existe fragilidade no fluxo de informação entre os responsáveis por registrar os óbitos e aqueles responsáveis em determinar intencionalidades das mortes. Por fim, utilizando abordagem bayesiana e dados de 2001 Luciana Tricai Cavallini¹⁷ sugere 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 ou o registro desta morte em categorias mal definidas, mortes por causas externas de intenção indeterminada (MVCI).

Quadro 1. Classificação dos óbitos quanto à intenção e 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

Em estudo com desenho ecológico, a investigação do número de homicídios ocultos utiliza informações registradas nas declarações de óbitos por causa externa, acessadas por meio

dos microdados do SIM, referentes ao período de 1996 a 2021 elaborando painel de dados com todas as Unidades da Federação (UFs) mais o Distrito Federal.

Quadro 1. Classificação dos óbitos quanto à intenção e instrumento.

Intencionalidade do óbito	Instrumento	Descrição da Categoria CID-10	Código CID-10
Indeterminada - MVCI	Envenenamento	Eventos (fatos) cuja intenção é indeterminada	Y10-Y19
	Enforcamento		Y20
	Afogamento		Y21
	PAF		Y22-Y24
	Impacto		Y25, Y30, Y31
	Veículo		Y32
	Fogo		Y26-Y27
	Perfurante		Y28
	Contundente		Y29
	Desconhecido		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: Autores, com base na CID-10.

Conforme descrito no Quadro 1, a investigação considera três tipos de morte por causa externa, em que o grupo homicídios totais (H^*) inclui as agressões e intervenções legais, o grupo de acidente\suicídio totais (AS^*) é definido como a união de mortes resultantes de acidente não natural ou suicídio e, por fim, o grupo MVCI. Acidentes\suicídios totais (AS^*) e homicídios totais (H^*) são apenas parcialmente observados, isto é, a porção não observada está classificada erroneamente como morte por causa externa de intenção indeterminada (MVCI). As equações 1,2 e 3 sintetizam o problema:

$$H^* = H + HO \quad (1)$$

$$AS^* = AS + ASO \quad (2)$$

$$MVCI = HO + ASO \quad (3)$$

Onde, HO e ASO são respectivamente, homicídios ocultos e acidentes\suicídios ocultos incorretamente registrados como morte por causa externa de intenção indeterminada (MVCI) e, H e AS são respectivamente, os homicídios e acidente\suicídios registrados no SIM, cuja intenção foi determinada pelo sistema de saúde. O objetivo do trabalho é identificar os homicídios ocultos (HO) registrados no SIM como morte por causa externa de intenção indeterminada MVCI. O não registro do homicídio no SIM ou registro deste em rubrica de mortes naturais implica considerar homicídios ocultos encontrados são limite inferior de incorreto registro de homicídios. Por sua vez, desde que a causa da morte por causa externa tenha sido determinada, supõe-se que a classificação está correta.

A identificação homicídios ocultos segue metodologia utilizada em diversos problemas de classificação¹⁸ e acontece em três etapas: otimização de hiperparâmetros (*model selection*), teste de generalização (*model assessment*) e ajuste do modelo final.

Na etapa de *model selection*, a base de dados composta por variável dependente binária da causa do óbito (homicídios (H) ou acidentes\suicídios (AS)) e variáveis previsoras foi dividida de forma aleatória estratificada em base de treinamento (70%) e base de teste (30%), de modo a manter em cada partição a proporção de homicídios e acidente\suicídio da base de dados original (*stratified random sampling*), por causa do desbalanceamento na base de treinamento, composta por 42,9% de homicídios, a base de treinamento foi reamostrada utilizando *Synthetic Minority Oversampling Technique* (SMOTE)¹⁹. Em seguida ocorre a otimização de hiperparâmetros utilizando *10-fold cross-validation*, assim, são ajustados sete algoritmos a base de treinamento (*logistic regression*, *penalized logistic regression* (*ridge*, *lasso* e *elastic net*), *decision tree*, *bagging decision trees* e *random forest*). São diversas as possíveis combinações de hiperparâmetros e nesta etapa busca – se identificar a combinação com melhor capacidade de previsão fora da amostra de cada algoritmo, por isso, a combinação de hiperparâmetros com maior média na área sob a curva ROC é selecionada.

A etapa de *model assessment* investiga possibilidade de *overfitting* e capacidade de generalização da combinação de hiperparâmetros selecionados na *model selection* ao realizar classificação da base de teste, isto é, a capacidade de

acertadamente elaborar classificação de observações inéditas, neste trabalho, a capacidade de o algoritmo identificar os homicídios ocultos. O algoritmo e 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 conjunto de dados (*i.e.* treinamento e teste), estabelecendo o modelo preditivo utilizado²⁰.

As variáveis preditivas são aquelas com informação ao longo de todo o período analisado, assim, os previsores são majoritariamente variáveis categóricas, 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*). Idade da vítima é a única variável contínua. Os *missing values* em idade da vítima são imputados através do método *k-nearest neighbors* ($k=5$) a partir das informações das mortes conhecidas e utilizando as demais variáveis com previsores da imputação. Por fim, a interpretação do modelo utiliza *SHAP Value* ao verificar a contribuição das variáveis predictoras na probabilidade estimada^{21,22}.

O método *Shapley Additive Explanations* (*SHAP*) explica a probabilidade estimada em cada observação ao atribuir a cada variável preditiva a contribuição percentual desta variável na probabilidade estimada, indicando o sentido e intensidade da associação entre o valor das variáveis predictoras e probabilidade estimada. Inspirado no *Shapley value*²³, 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:

$$\phi_j = \sum_{S \subseteq P \setminus \{j\}} \frac{|S|! (p - |S| - 1)!}{p!} (f_x(S \cup \{j\}) - f_x(S))$$

Ou seja, o *SHAP* da variável j realiza somatório das S coalizões de variáveis que não contém a j -ésima variável. A contribuição marginal da j -ésima variável em cada coalização será a diferença entre a probabilidade estimada incluindo

a j -ésima variável $f_x(S \cup \{j\})$ e sem a j -ésima variável $f_x(S)$, ponderado pela probabilidade da variável j contribuir na coalização S . Em resumo, o *SHAP* atribui importância a 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.

Estimação e interpretação dos modelos

A Tabela 1a apresenta a média das métricas de desempenho preditivo dos sete algoritmos ao aplicar *10-fold cross validation* na base de treinamento. Nesta etapa os algoritmos são otimizados, portanto, em cada algoritmo, apresentamos o desempenho da 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* (RF) foi aquele de melhor desempenho, superando os demais algoritmos nas métricas consideradas.

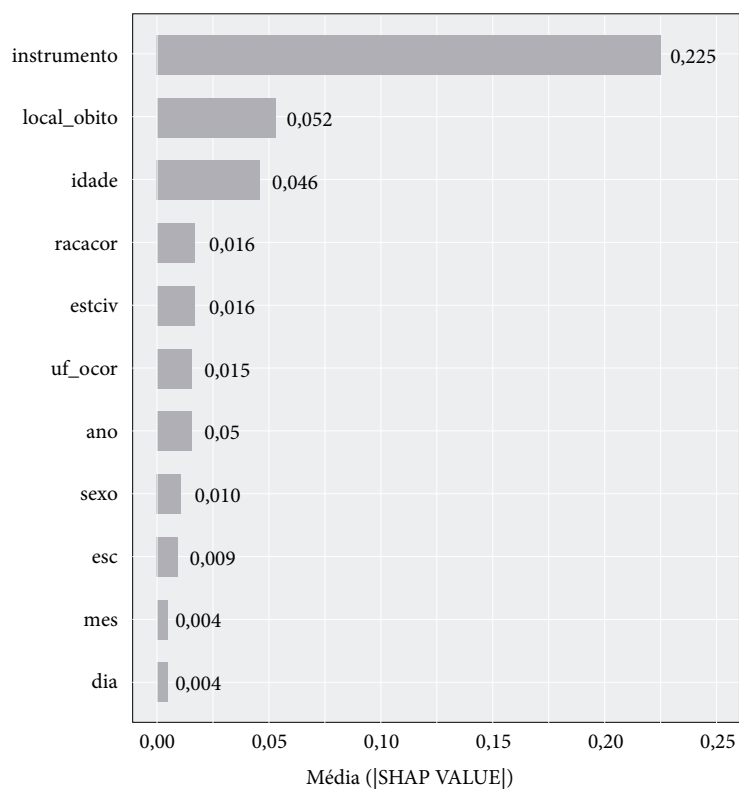
De acordo com as evidências apresentadas na Tabela 1b, similar ao processo de *model selection* na base de treinamento, nas previsões de observações inéditas da base de teste os algoritmos utilizados apresentam excelente desempenho preditivo. O bom desempenho na classificação de observações inéditas sugere adequada capacidade de generalização²⁴⁻²⁶ e ausência de *overfitting*. Na etapa de *model assessment* o *random forest* (RF), por pequena margem, apresentou o melhor desempenho nas métricas consideradas e, portanto, foi selecionado como modelo a ser utilizado na identificação dos homicídios ocultos.

A Figura 1 apresenta o *SHAP Summary Plot* da base de MVCI, isto é, a média dos valores absolutos dos *SHAP value* de cada variável em todas as probabilidades estimadas das MCVis serem homicídio oculto. O eixo y expõe, em ordem decrescente, a influência das variáveis preditivas, ou seja, ranking da importância relativa de cada variável na elaboração das probabilidades estimadas. O resultado indica 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.

Tabela 1. Métricas de performance no *cross validation* - *Model selection*.

A) Model Selection									
Modelos	Accuracy	Bal Accuracy	Precision	Sensitivity	Specificity	ROC AUC	F meas	J index	MCC
RF	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
B) Model Assessment									
Modelos	Accuracy	Bal Accuracy	Precision	Sensitivity	Specificity	ROC AUC	F meas	J index	MCC
RF	0,9625	0,9628	0,9486	0,9646	0,9610	0,9917	0,9566	0,9256	0,9237
Bagging Tree	0,9464	0,9610	0,9405	0,9593	0,9510	0,9901	0,9469	0,9175	0,9130
DTree	0,9590	0,9602	0,9380	0,9682	0,9522	0,9898	0,9529	0,9204	0,9170
Logit	0,9575	0,9584	0,9378	0,9646	0,9522	0,9899	0,9510	0,9168	0,9138
Lasso Logit	0,9574	0,9583	0,9377	0,9644	0,9522	0,9899	0,9509	0,9166	0,9136
Ridge Logit	0,9541	0,9547	0,9355	0,9588	0,9507	0,9875	0,9470	0,9094	0,9068
Elastic Net Logit	0,9574	0,9583	0,9377	0,9644	0,9522	0,9899	0,9509	0,9166	0,9136

Fonte: Autores.

**Figura 1.** SHAP *summary plot*.

Fonte: Autores.

Resultados e discussão

Durante o período compreendido entre 1996 e 2021, o sistema de saúde não identificou 128.567 homicídios. Portanto, os achados sugerem que 43,6% das MVCI registradas no Brasil são na realidade homicídios ocultos. Esse valor revela-se superior em 10 pontos percentuais a estudo similar investigando as MVCI na Rússia⁴. Na média anual o Estado foi incapaz de identificar 4.944 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 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 na literatura¹ a extensão da subnumeração de homicídios não é negligenciável.

A trajetória da proporção de homicídios ocultos incorretamente classificados na categoria MVCI, expõe três padrões de subnumeração. No intervalo entre 1996 e 2009, a média das proporções de 47,6% foi a maior em todo o período. Neste 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 MVCI atinge 37,1%, menor média da série, com valor mínimo de 35,4%. Ainda assim, nestes anos de menor subnumeração, ao menos um terço das MVCI são homicídios ocultos. No quadriênio final a incerteza sobre intencionalidade dos óbitos volta a aumentar e registra-se, em 2019, o ano de maior número absoluto de MVCI, entretanto, não ocorre expansão de mesma magnitude nos homicídios ocultos, embora a média anual tenha aumentado para 40,9% nesse mesmo quadriênio.

Ao evitar o viés estatístico decorrente da subnotificação de indicadores criminais, alguns estudos assumem premissa de que a proporção de subnotificação permanece constante ao longo do tempo nas áreas sob investigação²⁷. A evidência apresentada aponta substancial variabilidade na proporção de homicídios subenumerados, contrariando, ao menos no caso dos homicídios brasileiros, hipótese geralmente adotada na literatura.

Ao nível subnacional as cinco UFs com o maior número absoluto de homicídios registrados – São Paulo, Rio de Janeiro, Bahia, Minas Gerais e Pernambuco – são também aquelas com maior incidência de homicídios ocultos, abrangendo 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, São Paulo a de maior somatório em termos absolutos, sendo em vinte anos a UF de maior número absoluto de homicídios ocultos.

A Figura 2 apresenta o impacto dos homicídios ocultos em avaliações de criminalidade violenta, ao comparar as taxas por 100 mil habitantes de homicídio registrado e projetado, isto é, a soma de homicídios ocultos aos homicídios registrados, entre as dez UFs mais violentas de cada ano, além da taxa registrada e projetada do Brasil.

No caso das UFs, considerar os homicídios ocultos, aponta fictícia redução de criminalidade violenta em algumas UFs. Por exemplo, entre 1998 e 2009 o Rio de Janeiro, UF de maior mediana da taxa de homicídios ocultos, ascende de posição em nove oportunidades e, em 2019, somente após contabilizar os homicídios ocultos, o Rio de Janeiro aparece entre as dez UFs mais violentas. Neste ano são identificados no Rio de Janeiro 2.480 homicídios ocultos, valor correspondente a 70,0% dos homicídios registrados na UF. Em quatro ocasiões, Roraima também aparece entre as UFs mais violentas somente após a inclusão dos homicídios ocultos. Além disso, incluir os 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.

Exercício de similaridade entre as taxas registradas e projetadas encontra valores elevados da métrica raiz do erro quadrático médio (RMSE) e indica as UFs mais afetadas ao contabilizar os homicídios ocultos. Assim, seis UFs (Rio de Janeiro, Bahia, Sergipe, Rio Grande do Norte, Roraima e São Paulo) aparecem com RMSE superior a 4, resultado relevante dado as médias estaduais de homicídio registrado. Por exemplo, em São Paulo ao longo de todo período a taxa projetada é em média 17,7% superior à registrada ou 69,1% considerando somente o quadriênio final.

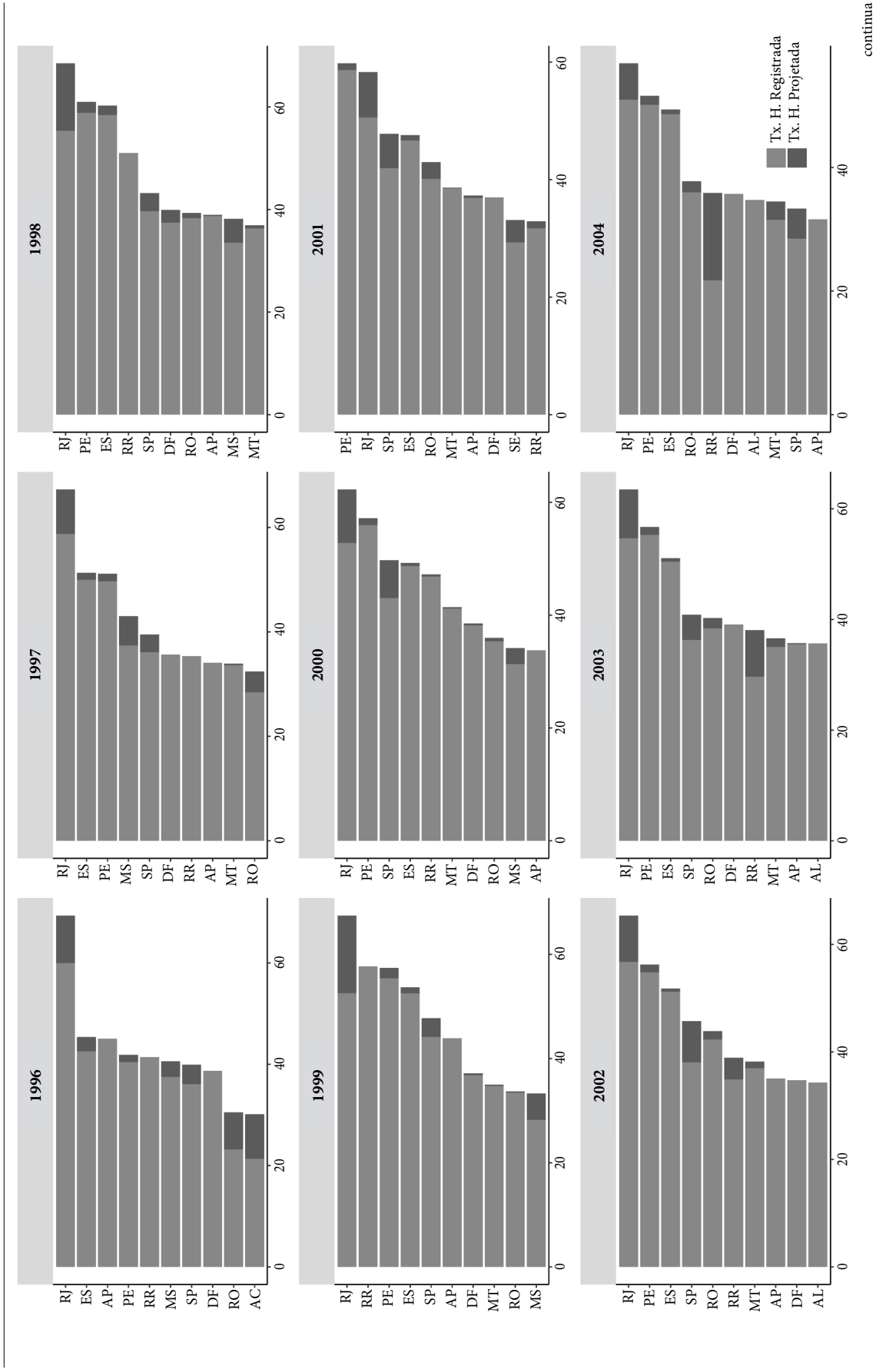


Figura 2. Taxa de Homicídio Projetado e Registrado, Brasil e UFs.

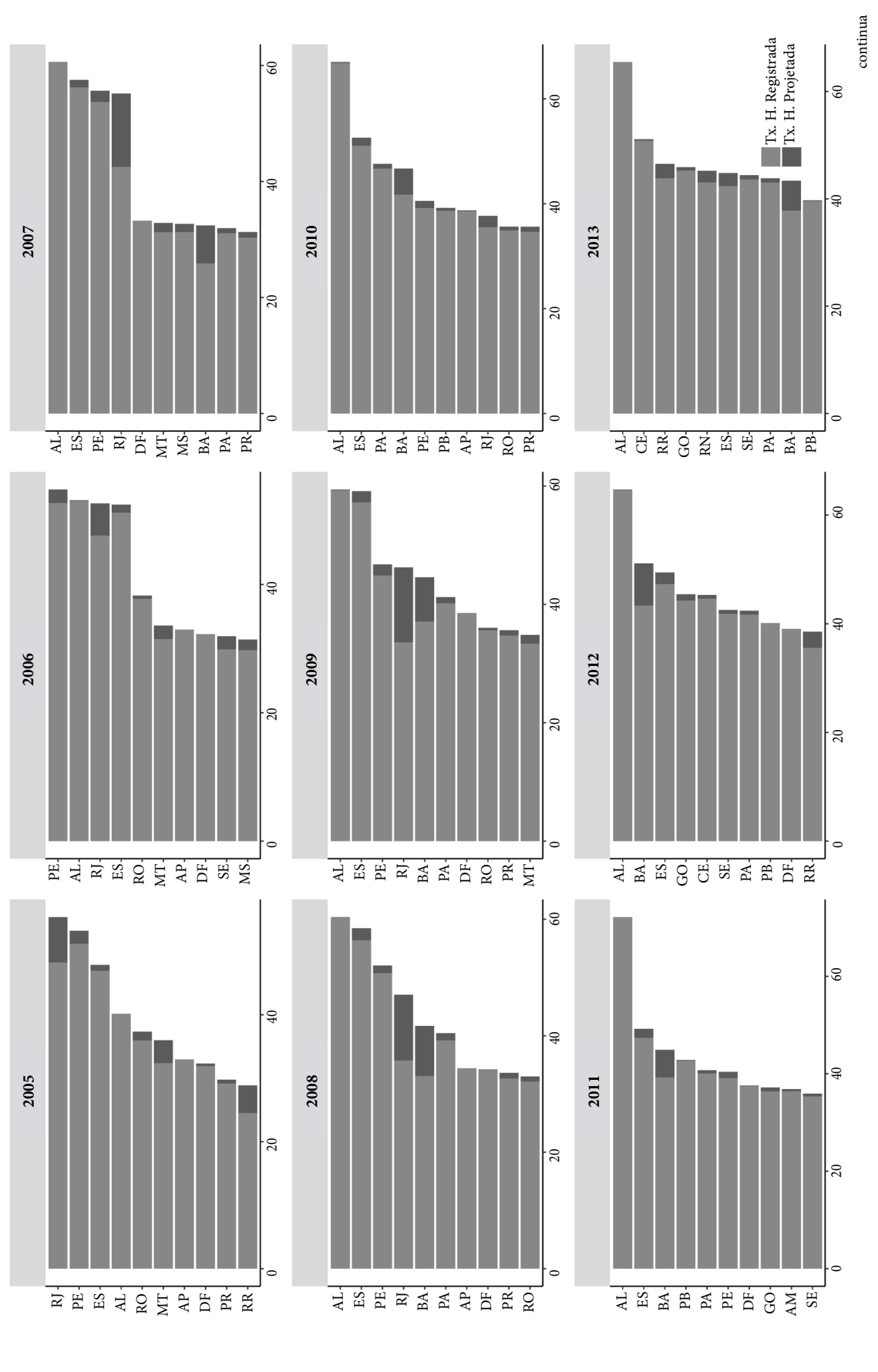
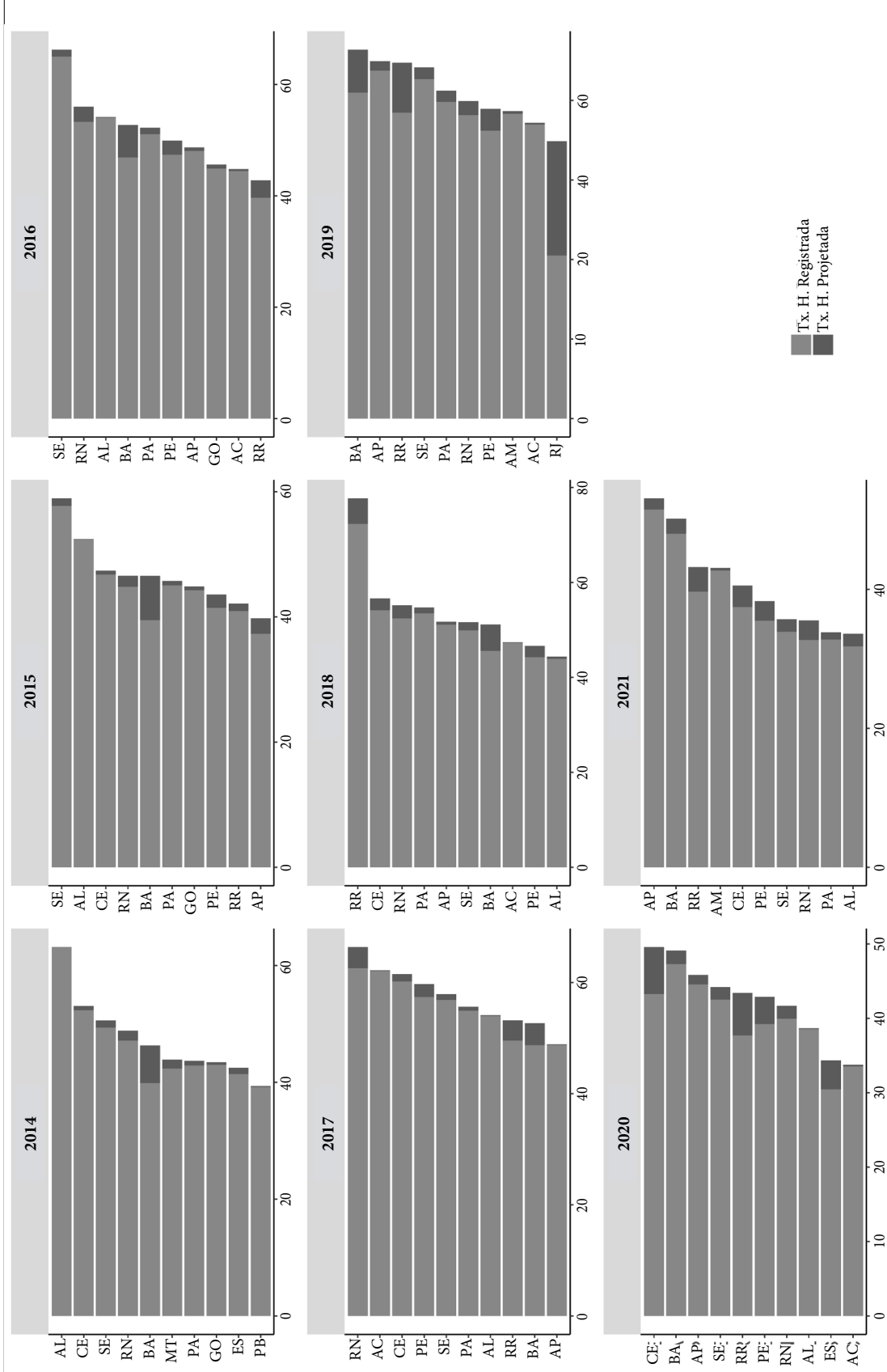


Figura 2. Taxa de Homicídio Projetado e Registrado, Brasil e UFs.



Este novo retrato da criminalidade violenta altera diagnósticos e avaliações de políticas de segurança. Por exemplo, no Rio de Janeiro a redução nos homicídios durante o período de implementação acelerada da Unidade de Polícia Pacificadora²⁸, 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 distância singular dentre as demais UFs e relevante em avaliações de política pública, a taxa registrada indica redução de 45,6%, enquanto a projetada sugere retração de 12,1%.

No agregado nacional a comparação entre taxa de homicídio registrado e projetado, utilizando métricas preditivas denota diferença relevante, por exemplo, 2,76 unidades no caso do RMSE (Raiz do erro quadrático médio). O elevado valor de erro, relativamente a média de homicídios registrados, sugere que a contabilização dos homicídios ocultos traz novo sentido às análises sobre a dinâmica de homicídios bra-

sileiros, 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 este indicador indica redução de 17,0%, diferença de cinco pontos percentuais, resultado relevante em avaliações de políticas públicas.

A diferença entre taxa de crescimento anual de homicídio registrado e projetado sintetiza o impacto da inclusão dos homicídios ocultos na série temporal de homicídios brasileira. De acordo com a Figura 3, as maiores distâncias entre as taxas de crescimento anual ocorrem no início e 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, achado que altera diagnósticos de segurança pública e revelador da importância em questionar a hipótese de subnotificação reduzida em homicídios. Como resultado das diferentes variações anuais, em média a taxa de homicídio projetada supera a taxa registrada em 10,0% e, no período analisado, o acumulado de homicídios projetados

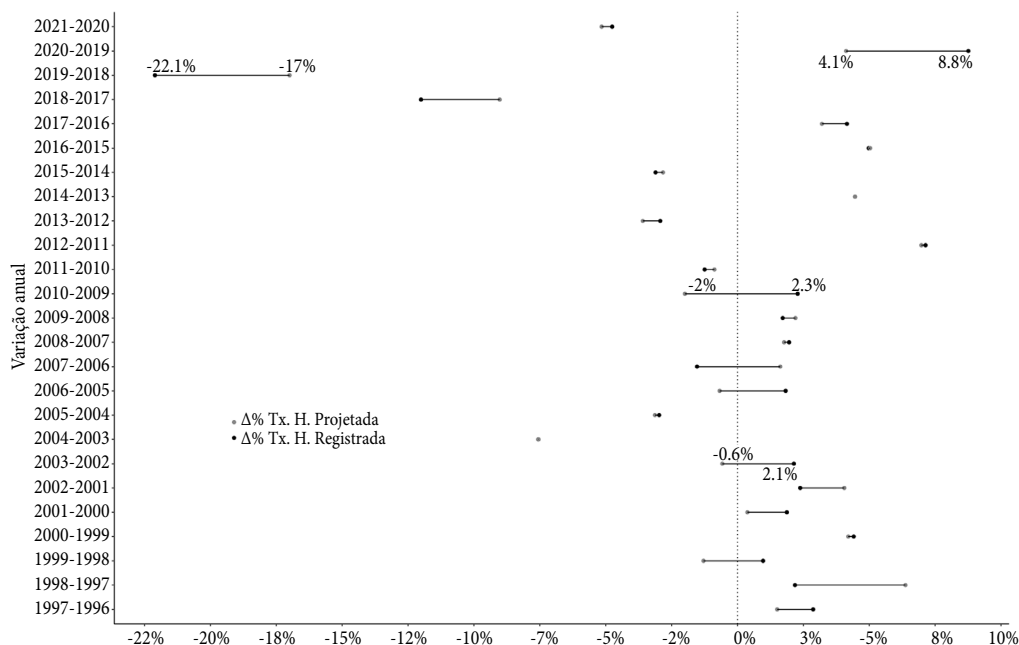


Figura 3. Taxa de crescimento anual homicídio registrada e projetada, Brasil.

Fonte: MS/SVS/Dados/SIM.

superou os registrados em 9,6%, totalizando 1.454.544 homicídios. Esses achados são evidência do imperativo de incluir os homicídios ocultos em investigações sobre criminalidade violenta. A ausência destes impossibilita diagnóstico capaz de guiar a focalização de intervenções nas UFs mais críticas ou caracterizar as populações mais vitimizadas, impedindo orientar políticas com objetivo de reduzir criminalidade violenta.

Conclusão

A utilização de homicídio como indicador dinâmica criminal é recomendada pela literatura e justificada por pesquisadores na hipótese, não verificada, de subnotificação reduzida^{1,7}. No Brasil o SIM é referência de informações sobre homicídios, sendo fonte de diversos estudos sobre criminalidade^{2,29}. Entretanto a ocorrência de elevada subnumeração de óbitos no SIM⁶, isto é, o não registro do óbito ou registro em categorias mal definidas e as evidências encontradas na literatura internacional^{4,14} e nacional¹⁵ sugerem que parcela dos óbitos por causa externa de intenção indeterminada (MVCI) são na realidade de homicídio oculto.

A partir dos microdados produzidos no preenchimento das declarações de óbito de mortes por causa externa registradas no SIM, entre os anos de 1996 e 2021, utilizando as características individuais e situacionais destas mortes, um conjunto de modelos de aprendizado supervisionado aprendeu as características dos homicídios e acidentes\suicídios e, em seguida, o modelo de melhor capacidade de generalização classificou as mortes por causa externa de intenção indeterminada em homicídios ou acidentes\suicídios, identificando limite inferior dos homicídios ocultos no SIM.

Desta forma, são identificados 128.567 homicídios ocultos, ou seja, 43,6% das mortes por causa externa de intenção indeterminada são na realidade homicídios ocultos. A quantidade de homicídios ocultos identificados é capaz de alterar diagnósticos sobre dinâmica criminal, avaliações de políticas públicas e invalidar pro-

cedimentos econométricos usualmente adotados na correção de subnotificação. Por exemplo, enquanto variações na distribuição temporal dos homicídios ocultos entre UFs e frequência da subnumeração, invalidam estratégias tradicionalmente utilizadas ao corrigir subnotificações no contexto de painel de dados²⁷ ou ao redistribuir homicídios³⁰, a identificação dos homicídios ocultos indica fictícia 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. Impactos demonstrados na subestimação do resultado das UPPs na trajetória de homicídios no Rio de Janeiro.

As evidências de subnumeração apresentadas por Soares Filho⁶ deveriam servir de alerta sobre os problemas na contabilidade de óbitos do SIM. Destes problemas, o elevado número de mortes por causa externa de intenção indeterminada é, *per se*, a principal objeção a utilizar homicídios registrados no SIM na suposição de subnotificação reduzida. A identificação de 128.567 homicídios ocultos sugere que o SIM não consegue registrar, em média, em cada ano do período analisado, um número de homicídios semelhante ao da Bahia, UF com o terceiro maior número absoluto de homicídios, valores estes que não podem ser negligenciados. Portanto, o expressivo número de homicídios ocultos contradiz a hipótese comumente adotada de subnotificação reduzida no registro de homicídios¹. A luz dos potenciais impactos negativos de toda sorte, ao utilizar o número de homicídios registrado no SIM é obrigatório estratégia capaz de reduzir o viés introduzido pelos homicídios ocultos. O método proposto, no entanto, não esgota as possibilidades de identificação de homicídios ocultos, apresenta capacidade de generalização positivamente correlacionada a qualidade no preenchimento das informações sobre óbitos com intencionalidade conhecida, portanto elevada incerteza sobre características do óbito pode afetar a capacidade preditiva do modelo, e não deve ser entendido como substituto da necessidade de aprimorar a qualidade das informações do SIM.

Colaboradores

DRC Cerqueira e GOA Lins contribuíram de forma equivalente em todas as etapas da pesquisa. Ambos participaram conjuntamente da concepção e delineamento do estudo, da revisão da literatura e da definição dos objetivos e hipóteses. Também atuaram em colaboração na coleta, organização e tratamento dos dados, bem como na realização das análises estatísticas e na interpretação dos resultados obtidos. A redação do manuscrito foi elaborada em conjunto, assim como as revisões críticas sucessivas, com contribuições intelectualmente relevantes de ambos. Por fim, os dois autores aprovaram a versão final a ser submetida para publicação.

Declaração de disponibilidade de dados

As fontes de dados utilizados na pesquisa estão indicadas no corpo do artigo.

Referências

1. Pinotti P. The Credibility Revolution in the Empirical Analysis of Crime. *Ital Econ J* 2020; 6(2):207-220.
2. Cerqueira D, Bueno S, Lima RS, Alves PP, Marques D, Lins GOA. *Atlas da Violência 2023* [Internet]. 2023 [acessado 2021 jun 10]. Disponível em: <https://www.ipea.gov.br/atlasviolencia/arquivos/artigos/9350-223443riatlasdaviolencia2023-final.pdf>.
3. World Health Organization (WHO). *International statistical classification of diseases and related health problems, 10th revision* [Internet]. 5ª ed. 2016 [acessado 2021 jun 10]. Available: <https://apps.who.int/iris/handle/10665/246208>.
4. Andreev E, Shkolnikov VM, Pridemore WA, Nikitina SY. A method for reclassifying cause of death in cases categorized as 'event of undetermined intent'. *Popul Health Metr* 2015; 13(1):23.
5. Värnik P, Sisask M, Värnik A, Yur'yev A, Kõlves K, Leppik L, Nemtsov A, Wasserman D. Massive increase in injury deaths of undetermined intent in ex-USSR Baltic and Slavic countries: Hidden suicides? *Scand J Public Health* 2010; 38(4):395-403.
6. Soares Filho AM, Cortez-Escalante JJ, França E. Review of deaths correction methods and quality dimensions of the underlying cause for accidents and violence in Brazil. *Cien Saude Colet* 2016; 21(12):3803-3818.
7. 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 BL, Zimmerman PR, editors. *Handbook on the Economics of Crime*. Cheltenham: Edward Elgar Publishing; 2010. p. 53-81.
8. Monteiro J, Caballero B. Crime e Violência. In: *Guia brasileiro de análise de dados: armadilhas & soluções*. 1ª ed. Brasília: Enap; 2021. p. 126-169.
9. Cerqueira D. Mortes violentas não esclarecidas e impunidade no Rio de Janeiro. *Econ Apl* 2012; 16(2):201-235.
10. Brasil. Lei nº 6.015, de 31 de dezembro de 1973. Dispõe sobre os registros públicos, e dá outras providências. *Diário Oficial da União*; 1973.
11. Brasil. Ministério da Saúde (MS). Conselho Federal de Medicina. Centro Brasileiro de Classificação de Doenças. *A declaração de óbito: documento necessário e importante*. 3ª ed. Brasília: MS; 2009.
12. Conselho Federal de Medicina (CFM). Resolução CFM nº 1.779/2005. Regulamenta a responsabilidade médica no fornecimento da Declaração de Óbito. *Diário Oficial da União*; 2005.
13. Ohberg A, Lonnqvist J. Suicides hidden among undetermined deaths. *Acta Psychiatr Scand* 1998; 98(3):214-218.
14. Santoro A. Recálculo de las tendencias de mortalidad por accidentes, suicidios y homicidios en Argentina, 1997-2018. *Rev Panam Salud Publica* 2021; 44:1.
15. Melo CM, Bevilacqua PD, Barletto M, França EB. Qualidade da informação sobre óbitos por causas externas em município de médio porte em Minas Gerais, Brasil. *Cad Saude Publica* 2014; 30(9):1999-2004.
16. Lopes AS, Passos VMA, Souza MFM, Cascão AM. 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. *Epidemiol Serv Saude* 2018; 27(4):e2018058.

17. Cavalini LT, Ponce de Leon ACM. Correção de sub-registros de óbitos e proporção de internações por causas mal definidas. *Rev Saude Publica* 2007; 41(1):85-93.
18. Nascimento CF, Santos HG, Batista AFM, Lay AAR, Duarte YAO, Chiavegatto Filho ADP. Cause-specific mortality prediction in older residents of São Paulo, Brazil: a machine learning approach. *Age Ageing* 2021; 50(5):1692-1698.
19. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res* 2002; 16:321-357.
20. Murphy KP. *Probabilistic Machine Learning: An introduction*. 1ª ed. Cambridge: Press, MIT; 2020.
21. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. Long Beach: NIPS 2017; 2017.
22. Aas K, Jullum M, Løland A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artif Intell* 2021; 298:103502.
23. James G, Witten D, Hastie T, Tibshirani R. Resampling Methods. In: *An Introduction to Statistical Learning*. New York: Springer; 2021. p. 197-223.
24. Hastie T, Tibshirani R, Friedman J. Model Assessment and Selection. In: *The Elements of Statistical Learning*. New York: Springer; 2009. p. 219-259.
25. Kuhn M, Johnson K. Over-Fitting and Model Tuning. In: *Applied Predictive Modeling*. 1ª ed. New York: Springer; 2013. p. 61-92.
26. Shapley LS. A Value for n-Person Games. In: *Contributions to the Theory of Games (AM-28), Volume II*. Princeton: Princeton University Press; 1953. p. 307-318.
27. Bianchi M, Buonanno P, Pinotti P. Do Immigrants Cause Crime? *J Eur Econ Assoc* 2012; 10(6):1318-1347.
28. Montes GC, Lins GO. Deterrence effects, socio-economic development, police revenge and homicides in Rio de Janeiro. *Int J Soc Econ* 2018; 45(10):1406-1423.
29. Cerqueira D, Soares RR. The Welfare Cost of Homicides in Brazil: Accounting for Heterogeneity in the Willingness to Pay for Mortality Reductions. *Health Econ* 2016; 25(3):259-276.
30. Garcia LP, Freitas LRS, Silva GDM, Höfelmann DA. Estimativas corrigidas de feminicídios no Brasil, 2009 a 2011. *Rev Panam Salud Publica* 2015; 37(4/5):251-257.

Artigo apresentado em 07/02/2024

Aprovado em 27/08/2025

Versão final apresentada em 29/08/2025

Editores-chefes: Maria Cecília de Souza Minayo, Romeu Gomes, Antônio Augusto Moura da Silva, Vânia de Matos Fonseca