

**FRAMEWORK PARA ANÁLISE DE DADOS CRIMINAIS MUNICIPAIS COM DADOS ADMINISTRATIVOS: UMA PROPOSTA METODOLÓGICA**

**FRAMEWORK FOR THE ANALYSIS OF MUNICIPAL CRIME DATA FROM ADMINISTRATIVE RECORDS: A METHODOLOGICAL PROPOSAL**

**FRAMEWORK PARA EL ANÁLISIS DE DATOS CRIMINALES MUNICIPALES CON REGISTROS ADMINISTRATIVOS: UNA PROPUESTA METODOLÓGICA**



10.56238/revgeov17n6-087

**André Felipe Gruber Bueno**

Doutorando em Gestão da Informação

Instituição: Universidade Federal do Paraná

E-mail: andre.gruber@ufpr.br

Orcid: 0000-0003-3494-4619

Lattes: <http://lattes.cnpq.br/1477189079040973>

**Luciano Heitor Gallegos Marin**

Doutor em Engenharia

Instituição: Université de Rennes I - França

E-mail: luciano.gallegos@ufpr.br

Orcid: <https://orcid.org/0000-0002-4331-6588>

Lattes: <http://lattes.cnpq.br/5686068199608499>

**RESUMO**

Secretarias estaduais de segurança pública brasileiras produzem rotineiramente registros administrativos de ocorrências criminais, mas carecem de protocolo analítico padronizado para transformar esses dados em diagnóstico territorial confiável. Este artigo propõe um framework em quatro etapas para análise quantitativa da criminalidade municipal a partir de dados administrativos: (1) diagnóstico distribucional e seleção do modelo de contagem adequado; (2) seleção de variáveis explicativas com fundamento teórico e validação estatística; (3) verificação cruzada por método preditivo independente; e (4) classificação municipal por risco absoluto e anomalia residual. O framework é apresentado como protocolo replicável, com atenção explícita às características estruturais dos dados administrativos criminais que condicionam as escolhas metodológicas em cada etapa. Discutem-se as condições de replicabilidade, as limitações internas ao framework e as extensões possíveis para contextos com dados de melhor qualidade.

**Palavras-chave:** Análise Criminal Quantitativa. Dados Administrativos. Gestão da Informação em Segurança Pública. Classificação Municipal de Risco.

**ABSTRACT**

Brazilian state public security secretariats routinely produce administrative records of criminal incidents, but lack a standardized analytical protocol to convert these data into reliable territorial diagnoses. This article proposes a four-step framework for the quantitative analysis of municipal crime



using administrative data: (1) distributional diagnosis and selection of the appropriate count model; (2) theoretically grounded variable selection with statistical validation; (3) cross-validation through an independent predictive method; and (4) municipal classification by absolute risk level and residual anomaly. The framework is presented as a replicable protocol, with explicit attention to the structural characteristics of administrative crime data that condition methodological choices at each step. Conditions for replicability, internal limitations, and possible extensions to higher-quality data contexts are discussed.

**Keywords:** Quantitative Crime Analysis. Administrative Data. Information Management in Public Security. Municipal Risk Classification.

## RESUMEN

Las secretarías estatales de seguridad pública brasileñas producen de manera rutinaria registros administrativos de incidentes criminales, pero carecen de un protocolo analítico estandarizado para convertir esos datos en diagnósticos territoriales confiables. Este artículo propone un framework de cuatro etapas para el análisis cuantitativo de la criminalidad municipal a partir de datos administrativos: (1) diagnóstico distribucional y selección del modelo de conteo adecuado; (2) selección de variables explicativas con fundamento teórico y validación estadística; (3) verificación cruzada mediante un método predictivo independiente; y (4) clasificación municipal por nivel absoluto de riesgo y anomalía residual. El framework se presenta como un protocolo replicable, con atención explícita a las características estructurales de los datos administrativos criminales que condicionan las decisiones metodológicas en cada etapa. Se discuten las condiciones de replicabilidad, las limitaciones internas del framework y las extensiones posibles para contextos con datos de mayor calidad.

**Palabras clave:** Análisis Criminal Cuantitativo. Datos Administrativos. Gestión de la Información en Seguridad Pública. Clasificación Municipal de Riesgo.



## 1 INTRODUÇÃO

Registros administrativos de ocorrências criminais são a principal fonte de dados disponível para análise territorial da segurança pública no Brasil. Secretarias estaduais acumulam milhões de registros por ano, com cobertura geográfica completa e desagregação temporal detalhada. Em teoria, esse volume de dados deveria suportar diagnósticos territoriais robustos. Na prática, a distância entre o dado disponível e o conhecimento utilizável pelo gestor é considerável, e parte dessa distância decorre da ausência de protocolos analíticos adequados às características específicas desses dados.

O problema não é novo na literatura de gestão da informação. Wang e Strong (1996) identificaram a adequação ao contexto de uso como dimensão central da qualidade informacional: uma informação é de qualidade quando o processo que a gera corresponde ao problema que se quer resolver. Dados criminais municipais têm características estruturais como sobredispersão severa nas contagens, subnotificação seletiva e heterogeneidade de escala entre municípios que tornam inadequados métodos desenvolvidos para outros contextos, mesmo quando esses métodos são tecnicamente sofisticados. Um modelo estatístico poderoso aplicado a dados com características mal diagnosticadas não produz informação melhor; produz informação incorreta com aparência de precisão (Choo, 2002).

A literatura internacional de criminologia quantitativa oferece métodos consolidados para análise de dados de contagem criminal (Cameron & Trivedi, 2013; Osgood, 2000), mas sua aplicação em secretarias estaduais brasileiras enfrenta dois obstáculos práticos. O primeiro é a dispersão do conhecimento metodológico: as escolhas que tornam uma análise válida estão documentadas em fontes técnicas que analistas de segurança pública raramente consultam sistematicamente. O segundo é a ausência de uma sequência lógica que organize essas escolhas: analistas que conhecem técnicas isoladas nem sempre sabem em que ordem aplicá-las ou como cada etapa condiciona as seguintes.

Este artigo responde a esse duplo obstáculo propondo um *framework* em quatro etapas para análise quantitativa da criminalidade municipal com dados administrativos. O *framework* não introduz métodos novos; organiza métodos estabelecidos em uma sequência lógica calibrada para as características específicas dos dados que secretarias estaduais brasileiras produzem. A proposta é explicitamente metodológica: o objetivo é oferecer um protocolo replicável, não um conjunto de achados substantivos sobre determinantes do crime.

O artigo está estruturado em quatro seções além desta introdução. A seção 2 caracteriza os dados administrativos criminais municipais e os problemas analíticos que eles apresentam. A seção 3 descreve as quatro etapas do *framework*. A seção 4 discute condições de replicabilidade, limitações internas e extensões possíveis.



## 2 DADOS ADMINISTRATIVOS CRIMINAIS: O QUE OS TORNA ANALITICAMENTE DIFÍCEIS

Antes de qualquer escolha metodológica, é necessário caracterizar os dados disponíveis. Dados administrativos criminais municipais brasileiros têm quatro características estruturais que condicionam todas as etapas da análise subsequente.

### 2.1 SOBREDISPERSÃO NAS CONTAGENS

A variável dependente em análises criminais municipais é quase sempre uma contagem: número de homicídios, furtos ou roubos em determinado município e período. O modelo estatístico mais simples para dados de contagem é a regressão de Poisson, que assume que a variância da variável dependente é igual à sua média.

Em dados criminais municipais, essa suposição é sistematicamente violada. A distribuição dos crimes entre municípios é extremamente assimétrica: alguns municípios concentram volumes muito acima do esperado pela média da distribuição, enquanto a maioria registra volumes baixos. O resultado é que a variância real é muito superior à média, fenômeno chamado sobredispersão (Cameron & Trivedi, 2013).

Ignorar a sobredispersão tem consequências diretas para a validade das análises. Modelos de Poisson aplicados a dados sobredispersos subestimam os erros-padrão dos coeficientes, produzindo testes de significância excessivamente otimistas e intervalos de confiança artificialmente estreitos (Gardner, Mulvey & Shaw, 1995).

O Teste da Razão de Verossimilhança entre um modelo de Poisson e um modelo Binomial Negativa é o procedimento padrão para detectar sobredispersão e determinar qual distribuição é adequada aos dados (Long, 1997).

### 2.2 SUBNOTIFICAÇÃO SELETIVA

Registros administrativos de ocorrências criminais dependem do comportamento de registro da população e da capacidade de captação das instituições. A taxa de subnotificação varia sistematicamente entre tipologias criminais e entre perfis municipais. Crimes contra o patrimônio com maior valor médio tendem a ser mais registrados porque as vítimas têm incentivo para comunicar à polícia, necessidade de boletim para seguros, por exemplo. Crimes violentos em municípios com menor presença institucional tendem a ser subregistrados porque as vítimas têm menor acesso às delegacias e menor confiança nas instituições (Cerqueira & Lobão, 2004).

A implicação analítica é que o número de ocorrências registradas não é equivalente ao número de ocorrências efetivas. Esse hiato não invalida a análise, mas requer cautela na interpretação: os modelos explicam a variação nos registros, não necessariamente a variação no crime real. Em



municípios pequenos com registro esporádico, ausência de registros pode refletir ausência de exposição estrutural ao risco ou ausência de capacidade de registro, e as duas interpretações têm implicações opostas para o gestor (Osgood, 2000).

### 2.3 HETEROGENEIDADE DE ESCALA ENTRE MUNICÍPIOS

Municípios brasileiros variam enormemente em população. Comparar contagens absolutas de crimes entre municípios com diferenças extremas de escala é metodologicamente incorreto: um município com 10 homicídios em uma população de 5.000 habitantes tem um problema qualitativamente distinto de um município com 10 homicídios em uma população de 500.000.

A solução técnica é transformar contagens em taxas por 100 mil habitantes ou, nos modelos de regressão, incluir o logaritmo da população como *offset*, um ajuste de exposição que não entra como preditor, mas controla a escala demográfica sem comprometer as propriedades estatísticas do modelo de contagem (Osgood, 2000).

O uso do *offset* é preferível ao cálculo prévio de taxas quando a variável dependente é analisada por modelos Binomial Negativa, porque preserva a natureza discreta da contagem original e trata corretamente municípios com populações muito pequenas, nos quais uma única ocorrência produz taxas por 100 mil habitantes numericamente extremas.

### 2.4 ESTRUTURA TEMPORAL E ESPACIAL NOS DADOS

Registros criminais têm estrutura temporal, o crime não se distribui uniformemente ao longo da semana ou do dia; e estrutura espacial: municípios geograficamente próximos tendem a apresentar taxas mais similares do que municípios distantes, seja por processos de difusão criminal, seja por características socioeconômicas compartilhadas (Anselin, 1995). Ignorar essas estruturas não as elimina dos dados: elas aparecem como autocorrelação residual que viola as suposições dos modelos e produz estimativas imprecisas.

A Teoria das Atividades Rotineiras oferece a fundamentação teórica para a dimensão temporal: a convergência entre infratores motivados, alvos adequados e ausência de guardiões varia sistematicamente com os ritmos sociais de trabalho, lazer e circulação (Cohen & Felson, 1979; Felson & Eckert, 2019).

Para a dimensão espacial, crimes patrimoniais oportunistas tendem a apresentar maior difusão entre municípios vizinhos do que crimes violentos, cuja raiz na desorganização social local os torna mais geograficamente concentrados (Weisburd, Groff & Yang, 2012).



### 3 O FRAMEWORK: QUATRO ETAPAS EM SEQUÊNCIA LÓGICA

O framework organiza as decisões analíticas em uma sequência em que cada etapa informa a seguinte. A ordem não é arbitrária: pular ou inverter etapas compromete a validade das conclusões. O Quadro 1 resume as etapas, seus objetivos e os critérios de decisão em cada uma.

Quadro 1 — Estrutura do Framework.

Etapa	Objetivo	Critério de decisão
1. Diagnóstico distribucional	Selecionar o modelo de contagem adequado	Razão variância/média; Teste da Razão de Verossimilhança Poisson vs. Binomial Negativa; coeficiente Pearson $\chi^2/gl$
2. Seleção de variáveis	Identificar correlatos estruturais com fundamento teórico	Teoria criminológica + AIC + Teste da Razão de Verossimilhança para remoções; VIF para multicolinearidade
3. Verificação cruzada	Validar a hierarquia de preditores por método independente	Verificar convergência GLM × Random Forest na hierarquia de importância; $R^2$ de validação cruzada
4. Classificação municipal	Produzir tipologia operacional de risco	Cruzamento de nível absoluto (taxa real vs. limiar) × anomalia residual (taxa real vs. taxa esperada)

Fonte: Elaborado pelos autores (2026).

#### 3.1 ETAPA 1: DIAGNÓSTICO DISTRIBUCIONAL

A primeira decisão analítica é a escolha do modelo de contagem. O ponto de partida é o diagnóstico da sobredispersão, que determina se a regressão de Poisson é adequada ou se a distribuição Binomial Negativa é necessária.

O procedimento padrão combina dois indicadores (Cameron & Trivedi, 2013). O primeiro é a razão variância/média da variável dependente: valores próximos de 1,0 indicam equidispersão compatível com Poisson; valores substancialmente superiores indicam sobredispersão.

O segundo é o Teste da Razão de Verossimilhança entre os dois modelos, que testa formalmente se o parâmetro de dispersão da Binomial Negativa é significativamente diferente de zero. Uma redução expressiva no AIC em favor da Binomial Negativa constitui evidência adicional. Após a estimação, o coeficiente de Pearson  $\chi^2/gl$  deve ser verificado: valores próximos de 1,0 indicam ajuste adequado sem sobredispersão residual relevante.

A Binomial Negativa tem um parâmetro de dispersão alpha que controla a variância adicional não capturada pela distribuição de Poisson. Esse parâmetro deve ser calibrado aos dados, não fixado arbitrariamente. A varredura sistemática do parâmetro com seleção pelo AIC, testando valores em uma grade, por exemplo, de 0,05 a 2,0 em incrementos de 0,05, é um procedimento reproduzível que evita a seleção *ad hoc* (Long, 1997).

Quando a variável dependente apresenta muitos zeros além do esperado pela Binomial Negativa, fenômeno chamado inflação de zeros, modelos *Zero-Inflated Negative Binomial* (ZINB)



devem ser considerados. O Teste de Vuong permite comparar formalmente a Binomial Negativa padrão com o ZINB e determinar qual distribuição se ajusta melhor aos dados (Vuong, 1989).

É importante distinguir zeros estruturais, municípios que genuinamente não registram o crime em análise por condições de baixa exposição, de zeros amostrais, municípios que poderiam registrar o crime, mas não o fizeram no período analisado. Zeros estruturais justificam a exclusão do subgrupo da modelagem, com documentação e justificativa explícitas; zeros amostrais demandam o ZINB.

### 3.2 ETAPA 2: SELEÇÃO DE VARIÁVEIS EXPLICATIVAS

A seleção de variáveis não deve ser conduzida apenas por critérios estatísticos. Em análise criminal municipal, as variáveis candidatas devem ser ancoradas em teorias criminológicas estabelecidas, de modo que os coeficientes estimados sejam interpretáveis e as hipóteses testáveis.

Duas tradições teóricas estruturam a maior parte das análises de criminalidade em nível municipal. A Teoria da Desorganização Social (Shaw & McKay, 1942; Sampson, Raudenbush & Earls, 1997) propõe que a violência interpessoal se concentra em territórios com menor capital humano, menor coesão institucional e menor eficácia coletiva. Indicadores de escolaridade, renda, desenvolvimento municipal e urbanização funcionam como proxies dessas condições estruturais, com associações esperadas negativas com homicídios e violência letal.

A Teoria das Atividades Rotineiras (Cohen & Felson, 1979) propõe um mecanismo oposto para crimes patrimoniais: o desenvolvimento econômico multiplica alvos acessíveis, eleva o poder de compra das populações e reduz a supervisão residencial, ampliando a oportunidade criminal independentemente do nível de coesão social local. Os mesmos indicadores de desenvolvimento operam, nessa perspectiva, com sinal positivo para crimes patrimoniais, configurando o paradoxo da riqueza (Felson & Eckert, 2019).

A implicação prática é que análises que tratam o crime como fenômeno unitário podem cancelar efeitos opostos e produzir resultados enganosos. A distinção por tipologia criminal é condição necessária para a validade dos modelos (Barros, Baggio & Baggio, 2020).

Com o conjunto candidato de variáveis definido teoricamente, a seleção estatística deve seguir um procedimento de refinamento em estágios orientado por critérios explícitos. O modelo saturado inclui todas as variáveis candidatas; remoções subsequentes são validadas pelo Teste da Razão de Verossimilhança para garantir que a variância explicada não se reduz significativamente (Cameron & Trivedi, 2013).

Multicolinearidade deve ser verificada pelo Fator de Inflação da Variância (VIF): valores acima de 5,0 indicam redundância que compromete a estimação dos coeficientes individuais e justifica a remoção de uma das variáveis correlacionadas (Hair et al., 2009).



Os coeficientes dos modelos Binomial Negativa devem ser reportados como Razões de Taxa de Incidência (IRR, do inglês Incidence Rate Ratio), obtidas pela exponenciação dos coeficientes na escala logarítmica. O IRR é interpretável diretamente como multiplicador da taxa esperada de ocorrências: um IRR de 0,85 para uma variável contínua indica que um aumento de uma unidade nessa variável está associado a uma redução de 15% na taxa esperada de crimes, mantidas constantes as demais variáveis. Intervalos de confiança de 95% devem acompanhar cada estimativa, calculados com erros-padrão robustos clusterizados por município quando a estrutura de dados inclui múltiplas observações por município (Cameron & Trivedi, 2013).

A inclusão de variáveis temporais (dia da semana ou período do dia) é recomendada quando a base de dados tem resolução temporal suficiente. Além de controlar variação temporal que, sem ajuste, inflacionaria os resíduos, essas variáveis documentam padrões de risco ao longo do tempo que têm interpretação teórica direta pela Teoria das Atividades Rotineiras (Cohen & Felson, 1979) e valor operacional para o gestor.

### 3.3 ETAPA 3: VERIFICAÇÃO CRUZADA POR MÉTODO INDEPENDENTE

A possível convergência entre dois métodos com lógicas distintas fortalece a confiança nos achados. Se um modelo paramétrico e um algoritmo não paramétrico identificam as mesmas variáveis como relevantes, na mesma hierarquia e no mesmo sentido de efeito, a evidência sobre os correlatos estruturais do crime é mais robusta do que se apoiada em um único método (Zhang et al., 2022).

O *Random Forest* (Breiman, 2001) é o algoritmo mais adequado para essa função de verificação em dados criminais municipais por três razões. Primeiro, opera sem suposições sobre a distribuição da variável dependente ou sobre a forma funcional das relações entre variáveis, tornando-o metodologicamente independente do GLM.

Segundo, produz uma medida de importância de variáveis, a redução média da impureza de Gini, que permite comparar sua hierarquia com a hierarquia de coeficientes do GLM.

Terceiro, modelos baseados em árvores têm demonstrado desempenho particularmente robusto em problemas envolvendo dados tabulares estruturados, frequentemente igualando ou superando abordagens alternativas mais complexas, especialmente quando há relações não lineares e interações entre variáveis (Grinsztajn, Oyallon & Varoquaux, 2022; Shwartz-Ziv & Armon, 2022).

A verificação cruzada deve documentar explicitamente dois elementos. O primeiro é a possível convergência hierárquica: as variáveis identificadas como mais importantes pelo *Random Forest* correspondem às variáveis com maior estatística *z* no GLM? A correspondência não precisa ser perfeita, mas divergências sistemáticas são um sinal de alerta que merece investigação.

O segundo é o desempenho preditivo do *Random Forest* em validação cruzada, que serve como indicador do grau em que os correlatos estruturais capturados pelos modelos explicam a variação



municipal no crime.  $R^2$  de validação cruzada muito baixo indica que os fatores mais determinantes da variação criminal não estão nas variáveis disponíveis, o que limita a interpretação dos coeficientes individuais (Breiman, 2001).

### 3.4 ETAPA 4: CLASSIFICAÇÃO MUNICIPAL POR RISCO E ANOMALIA RESIDUAL

A última etapa transforma os resultados dos modelos em um instrumento de gestão. A classificação municipal deve diferenciar municípios não apenas pelo nível absoluto de crime, mas pelo grau em que esse nível é explicado pelos correlatos estruturais ou constitui uma anomalia não observável pelos modelos.

A tipologia resulta do cruzamento de duas dimensões (Ratcliffe, 2011). A primeira é o nível absoluto de risco, calculado pela taxa real de crimes por 100 mil habitantes em relação a um limiar pré-definido. O limiar mais comum na literatura é a média mais um desvio-padrão da distribuição de taxas municipais, que identifica aproximadamente os 16% superiores da distribuição como municípios de alto risco absoluto (Long, 1997).

A segunda dimensão é a anomalia residual, diferença entre a taxa real observada e a taxa estruturalmente esperada pelo modelo. Um município com taxa real acima do esperado apresenta anomalia positiva; um município com taxa abaixo do esperado apresenta anomalia negativa.

O cruzamento das duas dimensões produz quatro categorias operacionais. Municípios com alto risco absoluto e anomalia positiva (taxa real acima do limiar e acima do esperado pelo modelo) são os casos de maior urgência gerencial, porque o nível de crime é tanto estruturalmente elevado quanto amplificado por fatores não capturados pelas variáveis disponíveis.

Municípios com alto risco absoluto e anomalia próxima de zero ou negativa têm crime elevado, mas estruturalmente explicado: as intervenções adequadas são as que atuam sobre os determinantes estruturais, não sobre fatores situacionais.

Municípios com risco baixo e anomalia positiva são casos latentes: o nível atual não é crítico, mas a pressão não observável sugere monitoramento intensificado.

Municípios com risco baixo e anomalia negativa (taxa abaixo do esperado dado o perfil estrutural) são os casos de resiliência: algo além das variáveis estruturais está suprimindo o crime, e identificar o que é apresenta valor para a formulação de políticas (Chainey & Ratcliffe, 2006).

Essa tipologia tem uma propriedade analítica central: ela separa o diagnóstico da intervenção. Municípios com o mesmo nível absoluto de crime podem demandar respostas distintas dependendo de onde estão na distribuição residual.

Um município Crítico-Anômalo demanda investigação dos fatores não observáveis, como redes de conflito local, presença de organizações criminosas, dinâmicas situacionais, que a análise estrutural não captura. Um município Crítico-Esperado demanda estratégias de médio e longo prazo

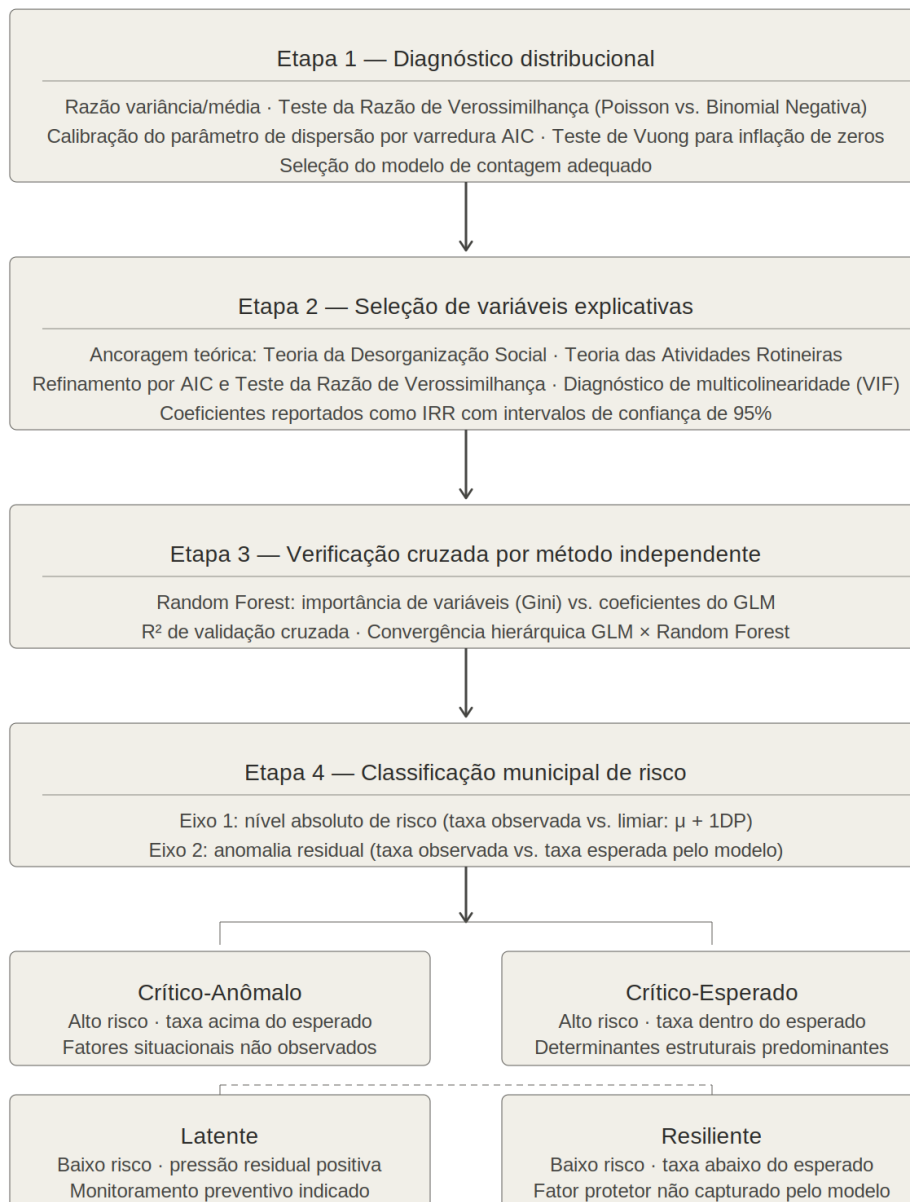


sobre desenvolvimento, educação e urbanização, que atuam sobre os determinantes identificados pelos modelos (Braga, Weisburd & Turchan, 2024; Sherman, 2013).

### 3.5 SÍNTESE DO FRAMEWORK

A Figura 1 representa esquematicamente a sequência lógica das quatro etapas descritas nas subseções anteriores. Cada etapa condiciona a seguinte: a escolha do modelo de contagem na Etapa 1 define a estrutura de estimação da Etapa 2; a hierarquia de variáveis validada na Etapa 2 é o objeto de verificação da Etapa 3; e os resíduos do modelo final alimentam o cruzamento dimensional da Etapa 4. A tipologia resultante é o produto analítico do framework como um todo.

Figura 1 — Sequência lógica do framework para análise de dados criminais municipais



Fonte: Elaborado pelos autores (2026).



A sequência não é reversível. Pular o diagnóstico distribucional da Etapa 1 e aplicar diretamente um modelo de Poisson, por exemplo, produz erros-padrão subestimados que invalidam os testes de significância da Etapa 2 e, por consequência, comprometem a hierarquia de importância verificada na Etapa 3. Da mesma forma, uma classificação municipal construída sem a verificação cruzada da Etapa 3 apoia-se em correlatos estruturais que podem refletir artefatos do modelo em vez de associações substantivas. A validade do *framework* depende da execução completa e ordenada das etapas.

## 4 CONDIÇÕES DE REPLICABILIDADE, LIMITAÇÕES E EXTENSÕES

### 4.1 CONDIÇÕES NECESSÁRIAS PARA REPLICAÇÃO

O *framework* depende de condições mínimas de qualidade e disponibilidade dos dados. A primeira é cobertura geográfica completa: o *framework* foi desenvolvido para análise do universo de municípios de uma unidade federativa, não para amostras. Análises amostrais introduzem viés de seleção que compromete a classificação residual, porque o limiar de risco absoluto é sensível à composição do grupo de referência.

A segunda condição é precedência temporal das variáveis explicativas em relação ao período criminal analisado. O uso de indicadores socioeconômicos do Censo Demográfico (IBGE) e de índices de desenvolvimento municipal para explicar ocorrências criminais registradas posteriormente atende a essa condição e fortalece a interpretação das associações, mesmo sem eliminar a possibilidade de variáveis confundidoras não observadas (Angrist & Pischke, 2009).

A terceira condição é resolução temporal mínima nos registros criminais. O *framework* opera de forma mais completa quando os registros incluem data e período do dia de cada ocorrência, permitindo a construção da estrutura de blocos que viabiliza os controles temporais da Etapa 2 e aumenta o poder estatístico dos modelos. Bases de dados que registram apenas o mês de cada ocorrência ainda permitem a aplicação das Etapas 1, 2 e 4, mas com modelos agregados no nível municipal-mensal em vez de municipal-diário-período.

A quarta condição é disponibilidade de *shapefile* dos limites municipais para construção da matriz de pesos espaciais utilizada no teste de autocorrelação residual. O IBGE disponibiliza esses arquivos gratuitamente para todos os Estados brasileiros, de modo que essa condição é satisfeita automaticamente em qualquer contexto estadual (IBGE, 2022).

### 4.2 LIMITAÇÕES INTERNAS DO *FRAMEWORK*

Quatro limitações merecem registro explícito. A primeira é a autocorrelação espacial residual. Mesmo após a inclusão de defasagem espacial e efeitos fixos mesorregionais, é possível que os resíduos apresentem agrupamento espacial significativo pelo Teste de Moran.



Isso indica a presença de processos locais que variáveis censitárias municipais não capturam. Modelos com componentes espaciais mais sofisticados, como o *Spatial Autoregressive Model* para dados de contagem, representam uma extensão metodológica relevante, embora sua estimação por máxima verossimilhança seja computacionalmente instável para distribuições Binomial Negativa (LeSage & Pace, 2009).

A segunda limitação é o desenho transversal. O *framework* produz associações estruturais entre variáveis socioeconômicas e crime em um corte de tempo, não estimativas causais. A precedência temporal das variáveis explicativas fortalece a plausibilidade causal, mas não a estabelece formalmente. Desenhos com variação temporal, como painel de municípios com múltiplos censos, ou análise de descontinuidades em políticas de segurança, permitiriam estimativas causais mais robustas ao custo de maior exigência de dados (Angrist & Pischke, 2009).

A terceira limitação é o escopo das variáveis censitárias. Indicadores de escolaridade, renda e urbanização capturam dimensões estruturais de médio e longo prazo, mas não fatores situacionais de curto prazo que podem explicar parte considerável da variação municipal no crime, em especial para tipologias como homicídios, cuja variação territorial é fortemente influenciada por dinâmicas de redes de conflito e disputas territoriais entre organizações criminosas (Cerqueira & Lobão, 2004). O baixo  $R^2$  de validação cruzada para homicídios em contextos com poucas variáveis explicativas é um indicador direto dessa limitação.

A quarta limitação é a sensibilidade do limiar de classificação. O critério média mais um desvio-padrão para definir municípios de alto risco absoluto é convencional, mas não universal. Em distribuições com assimetria muito acentuada, esse limiar pode ser sensível a municípios extremos. Análises de sensibilidade que testam limiares alternativos são recomendadas antes de usar a tipologia como base para decisões de alocação de recursos.

#### 4.3 EXTENSÕES PARA DADOS DE MELHOR QUALIDADE

O *framework* foi calibrado para as características típicas dos dados administrativos disponíveis nas secretarias estaduais brasileiras. Com dados de melhor qualidade ou maior resolução, três extensões são diretamente aplicáveis.

A primeira extensão é a análise em nível de bairro ou setor censitário, quando os registros criminais incluem georreferenciamento preciso da ocorrência. Nesse nível, variáveis de uso do solo, densidade comercial e configuração viária tornam-se preditores plausíveis ancorados na teoria dos padrões criminais (Brantingham & Brantingham, 2008), e a unidade de análise passa a capturar micro-heterogeneidades que o nível municipal suaviza.

A segunda extensão é a análise de séries temporais, quando há pelo menos três a cinco anos de dados consistentes. Modelos de painel com efeitos fixos municipais eliminam a influência de



características não observadas estáveis no tempo e permitem identificar o efeito de variações nos preditores sobre mudanças nas taxas criminais (Wooldridge, 2010).

A terceira extensão é a incorporação de variáveis operacionais de segurança pública: efetivo policial por município, número de delegacias, tempo médio de resposta a chamados. A literatura internacional é consistente em mostrar que recursos policiais têm efeito limitado sobre crime estrutural após controle por variáveis socioeconômicas (Lum & Koper, 2017; Sherman, 2013), mas a testabilidade dessa hipótese depende da disponibilidade de dados operacionais georeferenciados em nível municipal, que secretarias estaduais possuem, mas raramente disponibilizam de forma sistemática para análise.

## 5 CONCLUSÃO

Este artigo propôs um *framework* em quatro etapas para análise quantitativa da criminalidade municipal com dados administrativos: diagnóstico distribucional, seleção de variáveis com fundamento teórico, verificação cruzada por método independente e classificação municipal por risco absoluto e anomalia residual. Cada etapa responde a uma das características estruturais dos dados administrativos criminais que, quando ignoradas, comprometem a validade dos resultados.

O argumento central não é que existe um método ótimo para análise criminal municipal. É que a adequação entre o método e as características dos dados disponíveis é condição necessária para que a análise produza informação de qualidade, no sentido que Wang e Strong (1996) definem como adequação ao contexto de uso. Um modelo Binomial Negativa com seleção teórica de variáveis e verificação cruzada produz diagnóstico mais confiável do que um modelo tecnicamente sofisticado aplicado sem diagnóstico distribucional preliminar ou sem ancoragem teórica.

O valor prático do *framework* está na tipologia resultante da Etapa 4: a distinção entre municípios cujo crime é estruturalmente explicado e municípios onde o crime excede o estruturalmente esperado diferencia o tipo de intervenção necessária. Políticas estruturais de longo prazo, como educação, desenvolvimento, urbanização controlada, são adequadas para reduzir crime explicado por determinantes socioeconômicos. Intervenções situacionais e operacionais são necessárias onde o crime é anômalo em relação ao perfil estrutural (Braga, Weisburd & Turchan, 2024). Sem esse diagnóstico diferenciado, recursos de segurança pública tendem a ser alocados uniformemente entre perfis que demandam respostas distintas.

O *framework* é apresentado como protocolo replicável. Suas etapas são operacionalizáveis com ferramentas estatísticas disponíveis gratuitamente (R, Python), com dados publicamente acessíveis (Censo IBGE, índice FIRJAN) e com registros que qualquer secretaria estadual de segurança pública já produz.



**REFERÊNCIAS**

- ANGRIST, Joshua D.; PISCHKE, Jörn-Steffen. Mostly harmless econometrics: an empiricist's companion. Princeton: Princeton University Press, 2009.
- ANSELIN, Luc. Local indicators of spatial association — LISA. *Geographical Analysis*, v. 27, n. 2, p. 93-115, 1995. <https://doi.org/10.1111/j.1538-4632.1995.tb00338.x>
- BARROS, Pedro H. B.; BAGGIO, Henrique S.; BAGGIO, Ivã S. Os determinantes socioeconômicos da criminalidade no Brasil: o papel das externalidades e heterogeneidades espaciais. *Revista Brasileira de Segurança Pública*, v. 14, n. 2, p. 188-209, 2020. <https://doi.org/10.31060/rbsp.2020.v14.n2.1091>
- BRAGA, Anthony A.; WEISBURD, David; TURCHAN, Brandon. Focused deterrence strategies and crime control. *Criminology & Public Policy*, v. 23, n. 1, p. 1-36, 2024. <https://doi.org/10.1111/1745-9133.12671>
- BRANTINGHAM, Patricia J.; BRANTINGHAM, Paul L. Crime pattern theory. In: WORTLEY, Richard; MAZEROLLE, Lorraine (ed.). *Environmental criminology and crime analysis*. Cullompton: Willan, 2008. p. 78-93.
- BREIMAN, Leo. Random forests. *Machine Learning*, v. 45, n. 1, p. 5-32, 2001. <https://doi.org/10.1023/A:1010933404324>
- CAMERON, A. Colin; TRIVEDI, Pravin K. *Regression analysis of count data*. 2. ed. Cambridge: Cambridge University Press, 2013.
- CERQUEIRA, Daniel; LOBÃO, Waldir. Determinantes da criminalidade: arcabouços teóricos e resultados empíricos. *Dados*, v. 47, n. 2, p. 233-269, 2004. <https://doi.org/10.1590/S0011-52582004000200002>
- CHAINEDY, Spencer; RATCLIFFE, Jerry. *GIS and crime mapping*. Chichester: Wiley, 2006.
- CHOO, Chun Wei. *Information management for the intelligent organization: the art of scanning the environment*. 3. ed. Medford: Information Today, 2002.
- COHEN, Lawrence E.; FELSON, Marcus. Social change and crime rate trends: a routine activity approach. *American Sociological Review*, v. 44, n. 4, p. 588-608, 1979. <https://doi.org/10.2307/2094589>
- FELSON, Marcus; ECKERT, Mary A. *Crime and everyday life*. 5. ed. Thousand Oaks: Sage, 2019.
- GARDNER, William; MULVEY, Edward P.; SHAW, Esther C. Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin*, v. 118, n. 3, p. 392-404, 1995. <https://doi.org/10.1037/0033-2909.118.3.392>
- GRINSZTAJN, Léo; OYALLON, Edouard; VAROQUAUX, Gaël. Why tree-based models still outperform deep learning on tabular data. *Advances in Neural Information Processing Systems*, v. 35, p. 507-520, 2022.
- HAIR, Joseph F. et al. *Multivariate data analysis*. 7. ed. Upper Saddle River: Prentice Hall, 2009.



INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). Censo Demográfico 2022. Rio de Janeiro: IBGE, 2022. Disponível em: <https://sidra.ibge.gov.br>. Acesso em: 10 jan. 2025.

LESAGE, James P.; PACE, R. Kelley. Introduction to spatial econometrics. Boca Raton: CRC Press, 2009.

LONG, J. Scott. Regression models for categorical and limited dependent variables. Thousand Oaks: Sage, 1997.

LUM, Cynthia; KOPER, Christopher S. Evidence-based policing: translating research into practice. Oxford: Oxford University Press, 2017.

OSGOOD, D. Wayne. Poisson-based regression analysis of aggregate crime rates. *Journal of Quantitative Criminology*, v. 16, n. 1, p. 21-43, 2000. <https://doi.org/10.1023/A:1007521427059>

RATCLIFFE, Jerry H. Intelligence-led policing. 2. ed. Abingdon: Routledge, 2011.

SAMPSON, Robert J.; RAUDENBUSH, Stephen W.; EARLS, Felton. Neighborhoods and violent crime: a multilevel study of collective efficacy. *Science*, v. 277, n. 5328, p. 918-924, 1997. <https://doi.org/10.1126/science.277.5328.918>

SHAW, Clifford R.; McKAY, Henry D. Juvenile delinquency and urban areas. Chicago: University of Chicago Press, 1942.

SHERMAN, Lawrence W. The rise of evidence-based policing: targeting, testing, and tracking. *Crime and Justice*, v. 42, n. 1, p. 377-451, 2013. <https://doi.org/10.1086/670819>

SHWARTZ-ZIV, Ravid; ARMON, Amitai. Tabular data: deep learning is not all you need. *Information Fusion*, v. 81, p. 84-90, 2022. <https://doi.org/10.1016/j.inffus.2021.11.011>

VUONG, Quang H. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, v. 57, n. 2, p. 307-333, 1989. <https://doi.org/10.2307/1912557>

WANG, Richard Y.; STRONG, Diane M. Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems*, v. 12, n. 4, p. 5-33, 1996. <https://doi.org/10.1080/07421222.1996.11518099>

WEISBURD, David; GROFF, Elizabeth R.; YANG, Sue-Ming. The criminology of place: street segments and our understanding of the crime problem. Oxford: Oxford University Press, 2012.

WOOLDRIDGE, Jeffrey M. Econometric analysis of cross section and panel data. 2. ed. Cambridge: MIT Press, 2010.

ZHANG, Yonglong et al. A novel ensemble method for crime prediction. *IEEE Access*, v. 10, p. 74919-74930, 2022. <https://doi.org/10.1109/ACCESS.2022.3191103>

