Ambiente Urbano e Criminalidade: uma Abordagem de *Machine Learning* para o Caso da Cidade do Recife

Célio Henrique P. Belmiro - PIMES/UFPE
Raul da Mota Silveira Neto - PIMES/UFPE
Raydonal Ospina Martínez - PPGE/UFPE

Resumo. Este trabalho tem como objetivo verificar a associação entre características da vizinhança - setores censitários - e violência na cidade do Recife. Busca-se, por meio da aplicação de um conjunto de técnicas de regressão penalizadas, a construção de modelos parcimoniosos que auxiliem no entendimento da questão. Em especial, a utilização dos modelos propostos, *ridge*, *lasso* e *elastic net* são uma novidade no estudo da violência para o país, tanto pelo caráter inovador da abordagem quanto pela região geográfica de estudo - mesmo para abordagens tradicionais, poucos estudos se propõem a estudar a vizinhança com esse grau de desagregação no caso brasileiro. A cidade do Recife, objeto de estudo, situa-se como a mais antiga das atuais capitais brasileiras e uma das mais violentas capitais do país, motivação adicional no entendimento de fatores que podem estar associados as altas taxas de CVLI – Crime Violento Letal e Intencional – na cidade. Adicionalmente, dada a natureza heterogênea da distribuição das ocorrências no espaço, inclui-se na modelagem a presença de variáveis defasadas espaciais. Os resultados apontam para o incremento do desempenho preditivo de tais modelos em comparação com a abordagem tradicional, e possibilitam discutir, para a cidade do Recife, a associação entre a violência e uma série de características socioeconômicas, sociodemográficas e da vizinhança.

Palavras-chave: Violênca. Criminalidade. Regressão Penalizada.

Abstract. This work aims to verify the association between neighborhood characteristics - census tracts - and violence. Through the application of a set of penalizing regression techniques, the construction of models that allow the selection of a set of variables that help in the understanding of the question. In particular, the construction of the proposed models, ridge, lasso and elastic net are novelties in the study of violence for the country, both for the innovative nature of the approach and for the geographic region of study - even for traditional approaches, few papers propose to study the neighborhood with this degree of disaggregation in the Brazilian case. The city of Recife is the oldest Brazilian capital and one of the most violent capitals in the country, an additional motivation in the understanding of factors that may be associated with high CVLI - Intentional Letal Violent Crimes - rates. In addition, given the heterogeneous nature of the distribution of occurrences in space, the presence of spatial lagged variables is included in the modeling. The results point to the increase in the performance of such models, and allow to discuss, for the city of Recife, a series of socioeconomic characteristics of the neighborhood and its association with the violence.

Keywords: Violence. Criminality. Penalized Regression.

JEL: R19.
Área: 10 - Economia Regional e Urbana.

1 Introdução

Fenômeno marcante da vivência urbana, o crime, em suas mais diversas manifestações, está no cerne da atenção pública. Seu impacto direto no bem-estar e implicações para decisão de diferentes agentes tornam este complexo fenômeno objeto de estudo de diversos ramos das ciências sociais. Principalmente, economistas, criminologistas e sociólogos tem, ao longo dos anos, contribuído para o entendimento do fenômeno com uma série de estudos empíricos e desenvolvimentos teóricos que buscam identificar quais fatores estão associados às ocorrências de crimes nas cidades. Tais contribuições são sintetizadas através de estruturas analíticas, entre as quais, destacam-se a Teoria da desorganização social (Shaw e Mckay, 1942), a abordagem da Eficácia coletiva (Sampson e Raudenbush, 1999), a abordagem econômica tradicional do crime de Becker (1968) e a Teoria das Atividades Rotineiras (Osgood et al.,1996; Tremblay e Tremblay, 1998). Contribuições mais recentes ressaltam também a relevância de Interações sociais (Glaeser et al., 1996; Zenou, 2003) e da distância ao trabalho e amenidades locais (Verdier e Zenou, 2002), como situações catalisadoras para a ocorrência de crimes nas cidades. (GIBBONS, 2004) Tais diferentes perspectivas, em que pese suas pertinências, trazem, porém, grande desafio para o trabalho empírico que visa os condicionantes da criminalidade, uma vez que ampliam bastante o número de tais condicionantes que devem ser considerados nos trabalhos aplicados.

Na verdade, a profusão de diferentes teorias e abordagens para explicar a ocorrência de criminalidade urbana ao mesmo tempo que atesta a complexidade do fenômeno, também revela um desafio empírico: a arbitragem do pesquisador quanto às variáveis a considerar nos trabalhos aplicados, uma vez que nos modelos quantitativos econométricos tradicionais as fortes associações entre diferentes condicionantes dificulta a identificação das diferentes influências. Nos últimos anos, contudo, o desenvolvimento da capacidade computacional contribuiu tanto para o incremento na geração e obtenção de dados como no surgimento de novas técnicas de análise. O desenvolvimento das ferramentas na área de aprendizagem de máquina - *machine learning* - possibilitam analisar e tomar decisões sobre conjuntos de informações denominados *big data*. Tais técnicas, aplicadas em diversos campos da ciência, como em Economia, Psicologia, Física, Química e Engenharia, representam um subcampo da inteligência artifical cujo objetivo, como destacam Mehta et al. (2018), é desenvolver algoritmos capazes de aprender automaticamente com os dados.

De maneira geral, é possivel caracterizar um problema de aprendizagem supervisionado como próximo a um problema tradicional de regressão. Como Mehta et al. (2018) e Hastie, Tibshirani e Friedman (2008) argumentam, dado um determinado conjunto de dados X, desejamos obter um modelo $f(\beta)$, que é uma função dos parâmetros β, assumindo uma função custo $C(X, f(\beta))$, o que permite verificar quão bem o modelo proposto explica e performa sobre o conjunto de dados X. Os parâmetros β escolhidos são aqueles que minimizam a função custo adotada. Tratando-se de problemas envolvendo previsão, o primeiro passo é dividir o conjunto de observações em dois grupos distintos - mutualmente excludentes - denominados conjuntos de teste e validação. Assim, o ajuste do modelo é feito utilizando-se o conjunto teste e sua performance testada aplicando os coeficientes estimados para descrever os dados do conjunto de validação.

Como a adoção do modelo $f(\beta)$, por exemplo, é apenas uma hipótese para o processo gerador de dados, é extremamente comum que as abordagens envolvendo *machine learning* no estudo de determinado fenômeno envolvam múltiplos candidatos a modelos para o melhor ajuste dos parâmetros, como em Li (2015), Qiu (2015) e Vastrad et al., (2013). Surge então a necessidade de uma medida - ou medidas - que permitam comparar os resultados obtidos. Para tanto, neste trabalho, considera-se os critérios de seleção de *cross-validation*, AIC, BIC, o Erro Quadrado Médio de Previsão - MSE de previsão - e uma medida de apreensão da variabilidade do ajuste, o pseudo-R^2.

Na literatura internacional, a aplicabilidade de tais modelos no estudo da criminalidade não é novidade. Felson et al., (1999) utiliza regressão *ridge* para estudar, respectivamente, o conjunto de elementos contextuais relacionados a violência entre diferentes grupos americanos e, a relação entre crimes violentos e composição racial das cidades americanas. Já Wheeler (2008) compara associação

entre crimes violentos e o consumo de álcool e apreensão de drogas nas cidades de Houston e Columbus, utilizando um modelo de regressão geograficamente ponderada sem penalização e penalizado, com resultados que sinalizam o ganho de performance preditiva em relação aos modelos não penalizados.

No caso brasileiro, a violência, em especial o CVLI [1] - Crime Violento Letal Intencional - representa um importante desafio associado a implementação de políticas de segurança pública no conjunto das esferas governamentais, dada sua inter-relação com outras esferas da atividade criminosa, como o tráfico de drogas. O país figura, no cenário internacional, como um dos mais violentos do mundo, com uma taxa média de homicídios por 100.000 habitantes de 29.2 no período 2000 – 2013, atrás apenas da Colômbia, na América do Sul. Neste contexto, o estado de Pernambuco, dentre as unidades da federação, apresenta durante o período 2006 – 2016, trajetória singular entre os estados brasileiros, considerado em 2006 como segundo estado mais violento da nação – 52.6 homicídios por 100.000 habitantes, atrás apenas do estado de Alagoas (53.1). (ANDRADE, 2015; Atlas da Violência, 2018)

A despeito do resumido conjunto dos fatos descritos nos dois parágrafos acima, muito pouco se conhece sobre a distribuição intra-urbana das ocorrências de CVLI e o conjunto de fatores socioeconômicos, demográficos e espaciais associados. Assim, este trabalho busca contribuir para o entendimento da violência urbana brasileira em dois importantes aspectos. Primeiro, a partir da aplicação de um conjunto de técnicas de aprendizagem de máquina investiga a relação entre um grande conjunto de variáveis - são tratados mais de 50 condicionantes - que caracterizam os setores censitários – menor unidade geográfica de análise dos municípios, definida pelo IBGE com base em critérios populacionais – e a ocorrência de CVLI na cidade do Recife, durante o período 2008 – 2010. Segundo, como forma de comparar a capacidade destes modelos em apreender a variabilidade dos dados, são feitos cotejos entre os desempenhos empíricos utilizados e aqueles que utilizam técnicas mais tradicionais, como regressões MQO ou modelos que incorporam dependência espacial (especificações SDM e SLX, por exemplo).

Dado o caráter multidimensional da criminalidade, em especial dos homicídios – englobando um conjunto de características da vítima, do agressor e do ambiente em que se encontram – e do amplo conjunto de covariáveis para as quais buscamos associação com a ocorrência de CVLI nos setores censitários, a aplicação de um conjunto de técnicas que permitam a obtenção de coeficientes interpretáveis dados os usuais problemas associados a este tipo de análise – multicolinearidade e heterogeneidade – podem contribuir no entendimento da dispersão e dinâmica dos homicídios na região. Na verdade, poucos trabalhos, no caso brasileiro, se detém ao nível de desagregação geográfica utilizado. A utilização da localização das ocorrências e sua correspondência com os setores censitários da cidade, permite capturar de maneira mais precisa um conjunto mais amplo da heterogeneidade local, que se perde nas análises mais agregadas. Ainda, foge ao conhecimento dos autores deste trabalho, desenvolvimentos para o caso brasileiro que façam uso das técnicas de regressão penalizadas utilizadas no estudo da violência urbana.

Além desta introdução, o trabalho está organizado em mais 5 seções. A seguir, empreendemos uma breve revisão de literatura, investigando os principais desenvolvimentos teóricos que buscam compreender o fenômeno da criminalidade, bem como um conjunto de aplicações empíricas para o caso brasileiro, em especial para cidade do Recife. Na terceira seção, apresentamos o conjunto de modelos desenvolvidos e critérios de seleção e performance aplicados para então, na quarta seção, apresentarmos uma descrição e caracterização das bases de dados. Na seção 5, apresentamos os principais resultados referentes aos modelos de regressão tradicionais e os modelos penalizados *ridge*, *lasso*, *elastic net*. Por fim, na seção 6, discutimos sobre o conjunto de evidências gerados no entendimento da criminalidade na cidade e possíveis impactos na formulação de políticas de combate a violência e desenvolvimentos futuros na área.

[1] Por CVLI - Crime Violento Letal Intencional - atribuem-se os homicídios, latrocínios - roubo seguido de morte - e demais crimes violentos seguidos de morte, como estupro e lesão corporal. Neste estudo, utilizaremos em alguns momentos como sinônimo de CVLI, o termo homicídio, uma vez que representam cerca de 98% das ocorrências para a cidade do Recife, no período em questão.

2 Vizinhança e criminalidade: revisão de literatura

Nesta seção é apresentado e discutido um conjunto de desenvolvimento teórico e empírico na literatura internacional e brasileira que buscam compreender o fenômeno da violência urbana, mais especificamente dos homicídios, e que enfatizam a associação entre os elementos locais, fatores socioeconômicos e demográficos e a violência urbana, bem como da inter-relação entre a tríade agressor, vítima e local, por meio do estudo das motivações e características da vitimização. Por óbvias limitações de espaço, foge ao escopo deste trabalho, contudo, uma ampla revisão de literatura acerca do vasto conjunto de desenvolvimento teórico dentro das ciências sociais, em especial, a sociologia, economia e criminologia, que buscam compreender o fenômeno sobre uma miríade de abordagens. Assim, os trabalhos discutidos e apresentados a seguir compõem um núcleo substantivo de evidências, mas representam apenas parte do amplo conjunto que se desenvolveu, principalmente no fim do século XX e início do século XXI, em especial no caso brasileiro, onde toma forma no período um corpo de estudo da questão.

A literatura envolvendo vizinhança e criminalidade começa a tomar forma no decorrer do século XX com uma série de trabalhos, dentro das ciências sociais, que buscam entender a etiologia da violência urbana dentro um amplo conjunto de perspectivas sociais, econômicas e culturais. Deve-se reconhecer, como destaca Andrade (2015) que o crime é um fenômeno complexo, envolvendo as ações e motivações do indivíduo perpetrador, uma vítima e um local – espaço. Assim, na visão de Miethe e Regoeczi (2004), a construção de um quadro analítico amplo, unindo elementos que busquem apreender a motivação do agressor e o ambiente das ocorrências constitui-se fundamental na compreensão dos atos criminosos.

A abordagem econômica tradicional enfatiza a racionalidade dos agentes potenciais criminosos (Becker, 1968) e relaciona a atividade criminosa a tomada de decisão racional do indivíduo como uma análise custo-benefício entre os ganhos com a prática ilícita e fatores negativos associados, como probabilidade de detenção e rigidez da condenação e o custo de oportunidade da atividade criminosa, o salário no mercado formal. Uma série de trabalhos empíricos busca evidenciar a relação entre a tomada de decisão racional do indivíduo e fatores como urbanização, demografia, pobreza e desigualdade, como apontado em Cerqueira e Moura (2015) e Cerqueira e Lobão (2004). Já a teoria das atividades de rotina, discute a relação entre uma vítima em potencial, um agressor em potencial e uma tecnologia de proteção. Nesta perspectiva, é na ausência de recursos de proteção e exposição de potenciais vítimas a situações de maior probabilidade de ocorrência associadas a indivíduos perpetradores, motivados a cometer atividade criminosa, onde residem as ocorrências criminosas. Como destacam Cerqueira e Lobão (2004), baseando-se em pesquisas de vitimização, alguns desenvolvimentos empíricos tem evidenciado a relação entre o estilo de vida e a criminalidade, como em Osgood et al., (1996) e Tremblay e Tremblay (1998).

A teoria da desorganização social, como argumenta Sampson (1997), pode ser definida como a falta de capacidade de uma estrutura comunitária de empreender valores comuns em seus residentes e manter um efetivo de controle social sobre os mesmos. Trata-se, como bem descrevem Cerqueira e Lobão (2004) de uma abordagem sistemática em torno das comunidades locais e vizinhanças, incorporando as relações entre indivíduos pertencentes a comunidade que contribuam para o processo de socialização e fatores estruturais, socioeconômicos e demográficos que incorporem heterogeneidade étnica, mobilidade residencial, desagregação familiar e urbanização. Assim, a violência estaria associada às vizinhanças, no que se convencionou definir, em situação de desvantagem ou fragilidade social - *disadvantage* – e uma série de trabalhos empíricos, como em Sampson (2009 e 2012), Becker (2016), Cavalcanti et al., (2017 e 2018) e Light e Harris (2012) tem evidenciado tais relações.

Outro importante potencial determinante da criminalidade, em especial no caso brasileiro, refere-se à estrutura demográfica e de gênero da população. Um resultado bem documentado na literatura internacional, as motivações para o crime não são constantes no ciclo de vida do indivíduo, como destaca Cerqueira (2014, apud Thornberry, 1996). Legge (2008) e Hirschi e Gottfredson (1983), por exemplo, apresentam amplo conjunto de evidências que tomam a relação entre idade e crime como

fator presente em diversos grupos sociais e em diversos momentos do tempo. No caso, brasileiro, Mello e Schneider (2004), ao estudarem a dinâmica dos homicídios no estado de São Paulo, identificam o componente demográfico como um dos principais responsáveis pela redução observada nos índices de homicídios a partir dos anos 2000.

A despeito da extrema variação espacial nas ocorrências de crime dentro das cidades brasileiras, não são abundantes, contudo, os trabalhos que procuram apreender os condicionantes locais (intra-urbanos) da violência, baseando-se a maior parte dos trabalhos em comparações entre cidades ou estados, o que em certa medida reflete a maior disponibilidade de informações a partir destes níveis de agregação. Note-se, contudo, que para a cidade do Recife, foco desta pesquisa, Menezes et al. (2013) apresentam pertinente contribuição para o entendimento da distribuição dos homicídios a nível distrital - avaliando os bairros. De maneira geral, os resultados evidenciam a associação espacial existente entre regiões mais desiguais e o número de homicídios, gerando evidência empírica importante acerca da dimensão espacial na distribuição das ocorrências na cidade. E, mais recentemente, Cavalcanti et al. (2018) expandem a região de análise para o conjunto mais amplo de municípios que compõem a Região Metropolitana do Recife e encontram forte evidência empírica que relaciona o número de ocorrências não só de homicídios mas de outras categorias de crime - como roubo, estrupo e lesão corporal - ao consumo/tráfico de drogas nas regiões de análise, resultados que corroboram com evidências de De Mello (2015) para o estado de São Paulo.

Daudelin e Ratton (2017) discutem sobre a dinâmica dos mercados de drogas na cidade do Recife, associando a violência sistêmica inerente a estes mercados como elemento principal da relação existente entre tráfico, repressão policial e violência, especialmente nos mercados caracterizados como abertos e descobertos, mais expostos e, no geral, associados as transações entre grupos mais vulneráveis de indivíduos. Ainda, Andrade (2015) realiza uma excelente análise sobre os elementos sociológicos associados a ocorrência de CVLI no estado de Pernambuco, durante o período de 10 anos entre 2004 e 2014, identificando forte associação entre características sociodemográficas e as ocorrências.

Em que pese sua importância por explorar as importantes variações espaciais da criminalidade intra-urbana, note-se, porém, que estes poucos estudos enfrentam um difícil e significativo desafio. Por utilizarem métodos tradicionais de investigação - econometria tradicional ou espacial -, necessariamente devem, de alguma maneira, arbitrar quanto às escolhas das variáveis que são consideradas potenciais condicionantes da criminalidade, o que, além de limitar o escopo da análise, termina por privilegiar enfoques particulares da abordagem do crime. Ao adotar uma perspectiva de aprendizado com os dados -*machine learning* -, a análise exploratória deste texto segue caminho diverso; aquele de maior autonomia das variáveis em relação à própria relevância para o fenômeno da criminalidade.

3 Metodologia

Nesta seção iremos apresentar o conjunto de técnicas aplicadas no estudo dos fatores associados a ocorrência de CVLI nos setores censitários do Recife. Faremos uma breve caracterização das técnicas tradicionais de regressão com componente espacial e penalizadas. A abordagem tradicional, que toma forma nos modelos de mínimos quadrados ordinários - MQO - e de regressão logística - logit - são o ponto de partida das abordagens empíricas envolvendo *machine learning* por duas razões principais: primeiro, tais modelos tendem a evidenciar uma série de problemas nos dados, como multicolinearidade e heteroscedasticidade, questões que por si podem comprometer a qualidade dos coeficientes de tais estimativas. Segundo, é um objetivo comum das análises empreendidas comparar a performance do amplo conjunto de técnicas aplicadas frente ao desempenho da abordagem padrão.

Deve-se reconhecer ainda, que dado o caráter espacial do fenômeno estudado, com forte evidência empírica, inclusive para a cidade do Recife – como em Menezes et al., (2013) e Cavalcanti et al., (2018) - da associação espacial entre as ocorrências de CVLI e as unidades geográficas consideradas – bairros - a desconsideração de elementos das vizinhanças e sua associação com o crime pode gerar estimativas viesadas. Assim, como forma de incorporar características das vizinhanças e avaliar se a consideração, para os setores censitários, de tais variáveis contribui no entendimento do fenômeno, obtêm-se também

estimativas para o modelo SDM - *Spatial Durbin Model* – que leva em consideração a defasagem espacial das características dos vizinhos bem como da variável resposta nas vizinhanças. Finalmente, considera-se ainda nos casos dos modelos padrão e regularizados a introdução da defasagem espacial das covariáveis utilizadas – abordagem espacial denominada SLX, *Spatial Lagged X*.

3.1 Abordagens tradicionais: MQO e Regressão Logística

Na tradicional perspectiva de MQO, trata-se de obter a relação linear entre a variável dependente, y, sobre um conjunto X de covariáveis por meio da minimização da soma do quadrado dos erros – função custo associada ao modelo MQO padrão. O problema pode ser descrito pela expressão abaixo, onde y representa o vetor da variável resposta, X a matriz de covariáveis e ϵ é um vetor dos erros. Temos:

$$Y = X\beta + \epsilon \tag{1}$$

A estimação de β pode ser lida como resultado do problema de minimização da função custo $S(\beta)$, que pode ser descrita na forma $S(\beta) = ||y - X\beta||^2$, resultando no vetor de coeficientes estimados $\hat{\beta} = (X'X)^{-1}X'y$. Ainda, é possível mostrar que, estabelecidas as suposições (a) exogeneidade estrita: termo de erro tem média condicional zero, $E[\epsilon|X] = 0$; (b) independência dos regressões da matrix X, ou seja, matriz X tem posto-rank completo; (c) $var(\epsilon|X) = \sigma^2 I_n$, de onde se deriva a hipótese de homocedasticidade – variância constante e (d) normalidade dos erros, onde se assume que seguem uma distribuição normal com média zero e variância $\sigma^2 I$, na forma $\epsilon \sim N(0, \sigma^2 I)$, o estimador de MQO será não-viesado e possuirá variância mínima. (Greene, 2008)

A discussão das propriedades do modelo MQO padrão torna-se importante, no âmbito desta investigação, em virtude da possível estrutura de correlação entre as covariáveis, o que pode ocasionar o problema de multicolinearidade, resultado em estimativas sem significância e nas quais, dada a estrutura de dependências das observações, torna-se muito difícil sua identificação. Adicionalmente, se detectada a presença de autocorrelação espacial entre as observações, a violação da suposição de normalidade dos erros pode levar a estimativa de coeficientes viesados, uma vez que o caráter de associação das observações não está sendo levado em consideração no modelo.

Em casos que envolvem variável resposta dicotômica, como a ocorrência ou não de crime na região de análise, a modelagem ocorre por meio da regressão logística - ou logit. Aqui, y_i pode assumir os valores 0, com probabilidade π_i e 1, com probabilidade $1 - \pi_i$. Desejamos identificar as probabilidades π_i dependentes de um vetor de covariáveis x_i. [2] Uma forma inicial de obter tais probabilidades é adotar a função linear na forma $\pi_i = x_i'\beta$, onde β são os coeficientes da regressão. No entanto, π_i é um valor entre 0 e 1, enquanto x_i' pode assumir qualquer valor. Uma solução para esse problema é aplicar a função logit, como em Qiu (2015), para obter uma estimativa dos coeficientes. A função logit ou log das probabilidades pode ser obtida por:

$$\eta_i = logit(\pi_i) = log\frac{\pi_i}{1 - \pi_i} \tag{2}$$

Então, supondo um conjunto de n observações indepentes $y_1, , y_n$, binárias, pode-se constuir o modelo logit como uma função linear do preditor na forma $logit(\pi_i) = x_i'\beta$. O modelo, como destaca (Rodríguez, 2007) é linear generalizado com resposta binomial e função ligação logit. Os coeficientes não podem continuar a ser interpretados como na formulação de modelos lineares, agora cada coeficiente β_i representa uma mudança na função logit da probabilidade associada a uma mudança em β_i mantendo todos os demais preditores constantes. Tomando o exponencial da expressão descrita acima, pode-se isolar a probabilidade para a i-ésima observação, na forma:

[2] Uma ótima revisão do tópico pode ser vista em: Rodríguez, G. (2007). Lecture Notes on Generalized Linear Models. URL: http://data.princeton.edu/wws509/notes/ e em Qiu (2015), que apresenta, detalhadamente, o processo de estimação por máxima verossimilhança.

$$\frac{\pi_i}{1-\pi_i} = exp\{x_i'\beta\} \tag{3}$$

Que define um modelo para as probabilidades, tornando mais clara a interpretação dos parâmetros. Uma mudança no i-ésimo preditor, mantendo todas as demais variáveis constantes, pode-se ser obtida multiplicando as probabilidades por $exp\{x_i'\beta\}$. Ou seja, o coeficiente exponencial representa uma taxa de probabilidades. Por fim, a estimação do modelo será feita por máxima verossimilhança, de onde também é possível obter a função de perda logarítmica - *logaritmic loss function* - comumente usada em problemas de classificação para avaliar a perfomance dos modelos. A função de máxima verossimilhança e de perca logaritmica são dadas, respectivamente, por:

$$\begin{aligned} logL(\beta) &= \sum_{i=1}^{n}\{y_i log(\pi_i) + (1-y_i)log(1-\pi_i)\} \\ L &= -\frac{1}{n}\sum_{i=1}^{n}\{y_i log(\pi_i) + (1-y_i)log(1-\pi_i)\} \end{aligned} \tag{4}$$

Deve-se destacar ainda que, como na abordagem MQO padrão, embora tal modelo não pressuponha, necessariamente, uma estrutura de independência das covariáveis, a presença de multicolinearidade e autocorrelação pode comprometer a estimativa dos coeficientes, levando a resultados viesados e imprecisos.

3.2 Econometria Espacial: *Spatial Durbin Model* e *Spatial Lagged X Model*

Sob a suposição de um padrão de associação espacial entre a distribuição das ocorrências de CVLI na cidade do Recife, buscou-se apreender tal dinâmica por meio dos modelos espaciais SLX - *Spatial Lagged X* - e SDM - *Spatial Durbin Model*. O modelo SLX incorpora as características espaciais por meio de um vetor de parâmetros adicional, θ, que representa uma média dos k vizinhos mais próximos, a depender da escolha da matriz de defasagem espacial. Neste trabalho, optou-se por usar a matriz de pesos espaciais que considera como vizinho de determinado setor i, todos setores censitários, dando um peso maior ao vizinho mais próximo. O fator de ponderação escolhido, foi o inverso da distância euclidiana entre os centroides dos setores. Assim, utilizou-se a matriz inverso da distância como matriz de defasagem espacial. O modelo SLX pode ser descrito como:

$$y = X\beta + \theta WX + \epsilon \tag{5}$$

Onde X representa a matriz de covariáveis, características do setor censitário. E W representa a matriz de defasagem espacial. No entanto, pode-se ainda considerar que o CVLI, podendo ser enquadrado como uma das – ou a – manifestação mais extrema de violência contra o outro, pode estar associado as ocorrências nas regiões vizinhas. Assim, como forma de incorporar possíveis ganhos associados a estrutura de correlação entre a associação da variável dependente na vizinhança, estimaremos, adicionalmente ao modelo SLX, o modelo SDM - *Spatial Durbin Model*, que pode ser descrito por:

$$Y = \rho W_y + Z\gamma + \epsilon \tag{6}$$

Onde $Z = [X\ WX]$ e $\gamma = [\alpha\ \beta\ \theta]'$, observe que a estimação da defasagem da variável dependente, requer, adicionalmente, a estimação de um parâmetro adicional, ρ que mede a significância da defasagem da variável resposta dos vizinhos. Assim, é possível ainda que ρ não seja significante, indicando que a inclusão de tal componente espacial não contribui no entendimento do fenômeno. (Lesage e Pace, 2009) Assim, o modelo SDM será obtido, sem penalização da função custo, para a taxa de ocorrência no setor censitário e para a variável resposta binária – se houve ou não, durante o período de análise, ocorrência de CVLI no setor.

3.3 *Machine Learning*: modelos de regressão regularizados

Métodos de regressão regularizados, como regressões *ridge* e *lasso* foram criadas, nos últimos anos, principalmente como uma alternativa para reduzir os erros preditivos das estimações por mínimos quadrados ordinários (Vastrad et al., 2013). Como destaca Heldwing (2017), a aplicação de penalizadores à função custo MQO é uma forma de obter estimativas adicionando algum viés e se reduzindo a variância, o que melhora o poder preditivo do modelo, obtendo-se coeficientes mais interpretáveis e estáveis.

3.3.1 Regressão *ridge*

Inicialmente desenvolvida por Hoerl e Kennard (1970), a regressão ridge é uma forma de penalização quadrática da função custo no processo de estimação por mínimos quadrados que reduz e estabiliza os parâmetros do modelo por meio da adição, na função que minimiza o quadrado dos resíduos, de um termo penalizador ℓ_2. Na presença de preditores não correlacionados o estimador de MQO será estável, uma que vez a matriz $X'X$ será diagonal. No entanto, na existência de covariáveis altamente correlacionadas surge o problema de multicolinearidade e a matriz $X'X$ se aproxima de uma matriz singular. Formalmente, partindo-se da especificação padrão de um modelo de regressão linear $Y = X\beta + \epsilon$, temos:

$$\hat{\beta}_{ridge} = argmin_{\beta}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2\} \tag{7}$$

Na expressão acima, $\lambda > 0$ é o parâmetro de penalização, que controla a redução da magnitude dos parâmetros do modelo. Assim, quanto maior o valor de λ maior a penalização e mais próximo de zero serão os coeficientes. Seguindo a metodologia apresenta por Hastie, Tibshirani e Friedman (2009), uma forma equivalente de apresentar o problema é:

$$\begin{aligned} \hat{\beta}_{ridge} &= argmin_{\beta}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2\} \\ sujeito \quad a \quad &\sum_{j=1}^{p} \beta_j^2 \leq t \end{aligned} \tag{8}$$

A adição de termos positivos nos elementos diagonais da matriz $X'X$, resultando em uma nova matriz $(X'X + \lambda I)^{-1}$, garante que tal matriz será inversível para dado valor de λ. Note-se que na presença de ortogonalidade dos preditores as estimativas do modelo ridge serão apenas uma forma escalonada das estimativas obtidas por MQO, i.e, $\hat{\beta}_{ridge} = \hat{\beta}_{MQO}/(1 + \lambda)$. A solução para a regressão ridge é, assim, dada por:

$$\begin{aligned} \hat{\beta}_{ridge} &= (y - X\beta)'(y - X\beta) + \lambda\beta'\beta \\ \hat{\beta}_{ridge} &= (X'X + \lambda I)^{-1}X'y \end{aligned} \tag{9}$$

3.3.2 Regressão *lasso*

Similar ao modelo *ridge*, a regressão lasso consiste na aplicação de um penalizador de tratamento usando a norma ℓ_1 na função custo. Introduzido por Tibshirani (1996), a natureza do termo penalizador λ tende a produzir alguns coeficientes que são exatamente zero, propriedade que permite a aplicabilidade em problemas envolvendo seleção de variáveis na presença de multicolinearidade e grande número de preditores. Nesse sentido, a regressão performa tanto uma seleção de modelo (já que pode reduzir número de variáveis) quanto uma penalização dos coeficientes. Os coeficientes são obtidos a partir de um problema de minimização análogo:

$$\hat{\beta}_{lasso} = argmin_{\beta}\{\frac{1}{2}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j^2)^2 + \lambda\sum_{j=1}^{p} \mid \beta_j \mid\} \tag{10}$$

Seguindo a metodologia apresentada por Hastie, Tibshirani e Friedman (2009) pode-se apresentar o problema de minimização como descrito abaixo. Para valores baixos de t alguns coeficientes estarão próximos ou serão iguais a zero. Note-se que o penalizador ℓ_2 produz coeficientes iguais a zero, enquanto na regressão *ridge* o penalizador ℓ_1 reduz a magnitude dos parâmetros mas não zera determinados coeficientes. De (10), também aqui é possível fazer:

$$\begin{aligned} \hat{\beta}_{lasso} &= argmin_{\beta}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \\ sujeito \quad a \quad &\sum_{j=1}^{p} \mid \beta_j \mid \leq t \end{aligned} \tag{11}$$

3.3.3 Regressão *elastic net*

Uma terceira alternativa incorpora os penalizadores ℓ_1 e ℓ_2 como forma de incrementar a performance dos modelos lasso e ridge, constituindo-se, assim, o modelo rede elástica - ou *elastic net*. Em particular, como destaca Heldwing (2017), quando $n < p$ o modelo lasso pode apenas identificar um número n de coeficientes iguais a zero. Quando $n > p$ e os preditores são altamente correlacionados a regressão ridge tende a ter uma performance superior ao modelo lasso. Adicionalmente, na presença de 2 preditores - ou mais - altamente correlacionados o modelo lasso tende a selecionar um preditor e ignorar os demais correlacionados, i.e, assumirão valor zero. Desenvolvido por Zou e Hastie (2005), trata-se de resolver o problema de minimização:

$$\begin{aligned} \hat{\beta}_{enet} &= argmin_{\beta}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \\ sujeito \quad a \quad &\lambda\sum_{j=1}^{p}(\alpha\beta_j^2 + (1-\alpha) \mid \beta_j \mid) \end{aligned} \tag{12}$$

O parâmetro α controla a influência dos penalizadores. Se $\alpha = 0$ o algoritmo performa como uma regressão ridge, por outro lado, se $\alpha = 1$ o algoritmo performa como uma regressão lasso, para valores de α entre 0 e 1 tem-se um processo de penalização híbrido, que leva em consideração uma combinação dos dois penalizadores, como na expressão acima. O processo de minimização é realizado com a escolha de um conjunto de valores para α e então aplicando cross-validation para selecionar um λ ótimo para cada α. O par de parâmetros de regulariação $(\hat{\lambda}, \hat{\alpha})$ que minimiza o critério que cross-validation é então aplicado na estimação dos coeficientes (Heldwing, 2017).

3.4 Critérios de seleção

A performance geral de métodos de aprendizagem - *machine learning* - está relacionada, dentre outros elementos, com seu poder preditivo. Mais especificamente, com a capacidade preditiva em novos conjuntos de dados. A determinação de um conjunto de instrumentos para medir o desempenho dos modelos proposto é, assim, essencial. Foram adotados três critérios de seleção de modelos. O critério de Akaike - AIC - o critério Bayesiano - BIC - e o critério de *cross-validation*.

O critério de Akaike pode ser obtido por $AIC = -2log(\mathcal{L} \mid dados)) - K$ onde $\mathcal{L}$ é o valor da função de máxima verossimilhança. Já o critério Bayesiano, pode ser obtido por $BIC = -2ln\mathcal{L} + Klog(n)$. Por fim, o processo de *cross-validation* consiste em dividir o conjunto de dados em subconjuntos de K amostras, obter o ajuste do modelo para $K-1$ amostras e então o erro predito a K_{th} subamostra do

conjunto. Tipicamente, k assume o valor 5 ou 10. O modelo selecionado é aquele apresenta o menor erro amostral da estatística utilizada.

3.5 Critérios de Performance

Como critério de performance para comparação das diferentes abordagens, duas medidas foram aplicadas são utilizadas neste trabalho, o erro quadrado médio de previsão - MSE - e o Pseudo-R^2. O MSE representa uma medida da qualidade do estimador por meio da contabilização da média do quadrado da diferença entre os valores estimados e a variável resposta, podendo ser expresso como: $MSE(\hat{\theta}) = Var(\hat{\theta}) + vi\acute{e}s^2(\hat{\theta})$. Adicionalmente, neste trabalho, como medida de performance preditiva dos modelos propostos, estamos considerando o MSE de previsão, que consiste no erro quadrado médio obtido por meio do ajuste dos coeficientes obtidos no conjunto de dados de validação e sua respectiva comparação com a variável dependente estudada, dado por $MSE_{prev} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_{val} - y)^2$.

Já o pseudo-R^2 pode ser descrito como uma medida para comparar a qualidade do ajuste de diferentes modelos, seguindo a metodologia proposta por Florencio et al. (2012), que consiste no quadrado da correlação de Spearman entre a variável resposta e os valores preditos.

Por fim, para os modelos de resposta binária, é utilizada uma medida de performance que consiste na relação entre a taxa de verdadeiros positivos e de falsos positivos preditos - a área sob a curva - AUC. Um classificador perfeito apresentará uma curva ROC que se extende, verticalmente, de 0 a 1, produzindo um ponto em (0,1) e então horizontalmente, abrangendo a totalidade da área do quadrado unitário, assim, medidas da área sobre a curva mais próximas de 1 representam modelos com melhor perfomance, enquanto uma AUC de 0.5 indica completa aleatoriedade dos valores preditos.

4 Base de Dados

Três conjuntos distintos de informações são utilizados na construção de uma base de dados única, que possibilita estudar fatores sócio-econômicos e características do ambiente - como a presença de amenidades e desamenidades urbanas - associados à ocorrência de CVLI na cidade do Recife, durante o período de Julho de 2008 a Junho de 2010. O Note-se que o georreferenciamento dessas ocorrências permite associar cada ocorrência a um ponto específico no espaço e estabelecer uma micro vizinhança associada, no nosso caso, os setores censitários. A estratégia torna possível obter, por meio de informações do Censo e da Prefeitura do Recife, um conjunto bastante abrangente de características associadas ao ambiente constituindo-se e, até aqui, ainda não utilizadas nos estudos de criminalidade intra-urbana no Brasil. Esta seção apresenta uma descrição e caracterização das bases utilizadas neste estudo.

4.1 Secretária de Defesa Social - SDS/PE

As informações sobre a ocorrência de Crimes Violentos Letais Intencionais - CVLI - na cidade do Recife foram disponibilizadas pela Secretária de Defesa Social - SDS - e contém, para o período de Julho de 2008 a Junho de 2010, o georreferenciamento de 1674 ocorrências na cidade. Deste conjunto de dados é possível obter uma série característica das ocorrências, como gênero da vítima, idade aproximada e período do dia em que ocorreu o CVLI. Neste estudo, como estamos interessados nas características do ambiente e nossa unidade de observação são os setores censitários, a única informação relevante é a localização geográfica das ocorrências, que permitem associação aos demais fatores locais. Espera-se, também, que a consideração das características demográficas associadas a região de análise captem parte do efeito destas variáveis. A figura a seguir apresenta a distribuição das ocorrências por número e taxa.

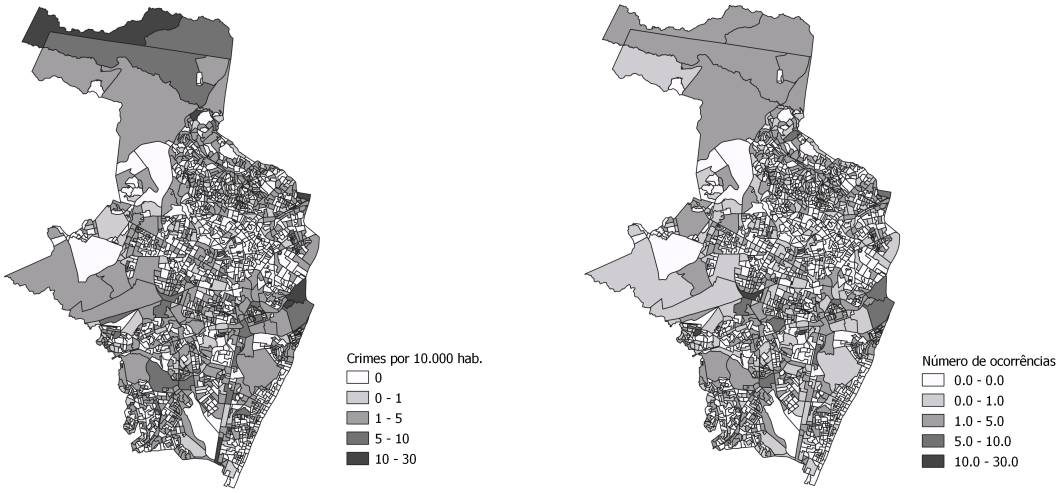

Figura 1: Distribuição da taxa de crimes (esquerda) e número de ocorrências (direita) por setor censitário na cidade do Recife. Fonte: elaboração própria.

4.2 Censo 2010

O conjunto de observações que permite obter as características socioeconômicas das regiões de análise foi obtido a partir dos dados do Censo Demográfico para o ano de 2010. Disponibilizado pelo IBGE - Instituto Brasileiro de Geográfica e Estatística - é descrito como: a principal fonte de referência para o conhecimento das condições de vida da população em todos os municípios do país e em seus recortes territoriais internos. As informações foram obtidas por setor censitário, a menor subdivisão geográfica dos municípios, definida pelo IBGE com base em critérios populacionais, respeitando a divisão político-administrativa brasileira.

4.3 Prefeitura da Cidade do Recife

A presença de amenidades e desamenidades[3] urbanas tem papel fundamental na caracterização da cidade. Existem dois ganhos explícitos na sua consideração: (a) por se tratar de um corte transversal, com informações disponível apenas para o ano de 2010, a consideração das amenidades urbanas capta parte do efeito fixo associado aos setores censitários e (b) de forma complementar, dada sua integração ao espaço, tem papel importante na decisão individual de moradia, bem como do poder administrativo na concepção de políticas públicas na cidade. Assim, consideramos as distâncias euclidianas das fronteiras dos setores censitários à praia de Boa Viagem, ao Rio Capibaribe, ao conjunto de parques e praças, às 53 avenidas arteriais da cidade, à rodovia BR-101 (única BR que cruza a cidade) e ao centro da cidade - região administrativa que concentra maior parte do emprego (Rodrigues et al. 2018). Em adição, uma vez que representam óbvias situações de presença do poder público, consideramos a presença de delegacias e batalhões de polícia, policlínicas, hospitais, escolas da rede municipal de ensino, além da condição do setor censitário estar ou não situado na fronteira da cidade (é contíguo a outro setor não pertencente ao município do Recife).

Uma descrição de todo o conjunto de variáveis é apresentado na Tabela 1. No total, são consideradas 53 variáveis potencialmente associadas à ocorrência de CVLI na Cidade do Recife

[3]Estamos considerando como amenidades e desamenidades urbanas um conjunto de características da cidade do Recife que, de acordo com evidências empíricas, como em Seabra (2013) e Rodrigues et al., (2018) tem impacto na decisão locacional dos indivíduos e no preço dos imóveis da região.

que capturam diferentes dimensões da vida social das vizinhanças. Destas, 22 variáveis dizem respeito às características demográficas e sociais das vizinhanças através de informações de seus domicílios, incluindo desde tradicionais informações sobre condições de renda como composição etária e informações sobre a escolaridade. Além destas, há mais 17 variáveis que informam sobre as condições de infraestrutura domiciliar dos setores censitários, como, por exemplo, acesso a saneamento e iluminação pública. Finalmente, o conjunto é completado com mais 14 variáveis que informam sobre as condições de localização dos setor censitário em relação a diferentes aspectos e equipamentos públicos da cidade, entres os quais, por exemplo, presença de escolas, delegacias e posição com respeito às amenidades urbanas.

Tabela 1: Descrição das variáveis

Variáveis	Descrição
num_ocorrencias	Número de ocorrências CVLI
txcrime	Taxa de ocorrências por 10.000 habitantes
rend_mensal_media	Valor do rendimento médio mensal das pessoas de 10 anos ou mais de idade
resp_mulher	% de responsáveis pelo domicílio do sexo feminino
mulher	% de indivíduos do sexo feminino
idade0_14	% de indivíduos entre 0 e 14 anos de idade
idade15_25	% de indivíduos entre 15 e 25 anos de idade
idade26_35	% de indivíduos entre 26 e 35 anos de idade
idade36_45	% de indivíduos entre 36 e 45 anos de idade
idade46_55	% de indivíduos entre 46 e 55 anos de idade
idade55mais	% de indivíduos com 55 ou mais anos de idade
alfabetizados_15mais	% de indivíduos alfabetizados com 15 ou mais anos de idade
negros_pardos	% de indivíduos negros e pardos
renda_dom_percmed	Renda domicilar per capita média
renda_dom_pc05	% de domicílios com renda nominal mensal per capita 0.5 salário mínimo
renda_dom_pc025	% de domicílios com renda nominal mensal per capita 0.25 salário mínimo
renda_chefe05	% de chefes de domicílio com renda nominal mensal de até 0.5 salário mínimo
chefe_semrenda	% de chefes de domicílio sem rendimento nominal mensal
soma_renda10	Rendimento médio per capita dos indivíduos com 10 ou mais de idade
dom_abastagua	% de domicílios com abastecimento de água
dom_saneamento	% de domicílios que tinham banheiro ou sanitário e esgotamento sanitário
dom_eletricidade	% de domicílios com acesso a luz elétrica
dom_coletalixo	% de domicílios com coleta de lixo
dom_10salarios	% de domicílios com renda mensal per capita acima de 10 salários mínimos
dom_moradequada	% de domicílios com infraestrutura completa* (moradia adequada)
dom_pavimentacao	% de domicílios com pavimentação
dom_iluminacao	% de domicílios com iluminação pública
dom_calcamento	% de domicílios com calçada
dom_esgotoaberto	% de domicílios com esgoto ao céu aberto
dom_arborizacao	% de domicílios com arborização
dom_lixoacumulado	% de domicílios com lixo acumulado
morador_pordom	Número médio de morados por domicílio
dom_3oumaisbanheiro	% de domicílios com 3 ou mais banheiros para uso exclusivo dos moradores
dom_semhomem	% de domicílios sem moradores do sexo masculino
dom_alugados	% de domicílios alugados
casas	% de domicílios que são casa (vs.apartamentos)
resp_alfabetizado	% de indivíduos responsáveis pelo domicílio alfabetizados
homem_15maisalfa	% de homens com 15 anos ou mais de idade alfabetizados
neg_par_15maisalfa	% de negros e pardos com 15 ou mais alfabetizados

Tabela 1 – Descrição das variáveis (continuação)

Variáveis	Descrição
pessoa_sregistro	% de indivíduos de até 10 aos de idade sem registro de nascimento
pessoa_srenda	% de indivíduos que possuem renda
areakm2	Área do setor censitário por km
denskm2	Densidade populacional do setor censitário por km
dpolicia	Se há no setor censitário delegacias de polícia ou batalhões de polícia
dsaude	Se há no setor censitário hospitais públicos e policlínicas
descola	Se há no setor censitário presença de escolas da rede municipal
dtiposetor	Tipo do setor censitário (normal ou subnormal)
dfronteira	Se o setor censitário encontra-se na fronteira do município
dist_avenidaskm	Distância em km da avenida mais próxima
dist_cbdkm	Distância em km da centro de negócios da cidade
dist_parqpracakm	Distância em km do parque ou praça mais próximo
dist_praiakm	Distância em km da praia de Boa Viagem
dist_riokm	Distância em km do ponto mais próximo do Rio Capibaribe
dist_rodoviakm	Distância em km do ponto mais próximo da rodovia federal (BR-101)
dist_metrokm	Distância em km da estação de metrô mais próxima

Fonte: elaboração própria.

5 Resultados

Os resultados da pesquisa são apresentados e discutidos a partir da seguinte estruturação. Na primeira subseção são apresentados os critérios de seleção dos modelos tradicionais (incluindo aqueles com dependência espacial) e penalizadores (*machine learning*). Estimativas para os coeficientes das variáveis para os modelos selecionados são apresentadas na segunda subseção, que também resume a importância dos grupos de variáveis consideradas. Em todos os casos, são considerados especificações com a variável dependente em termos de taxa de homicídios (ocorrência por 100.000) e com a mesma como indicador binário de ocorrência de homicídio (modelos logit).

5.1 Seleção dos modelos

5.2 Modelos tradicionais e com dependência espacial

De início, a suposição inicial de dependência espacial das observações pode ser testada de modo a verificar se, em conformidade com a evidência empírica já existente para os bairros, em Menezes et al. (2013), evidencia-se algum padrão de associação espacial entre as ocorrências para os setores censitários. Para tanto, aplicaremos duas medidas amplamente utilizadas na literatura que permitem verificar e testar a existência de algum padrão subsequente de associação espacial. São elas o I de Moran e o LISA - Indicadores Locais de Autocorrelação Espacial. A tabela 2, abaixo, apresenta a estatística de I de Moran para a taxa e o número de ocorrências. É possível então perceber, a principio, que embora o índice seja baixo, parece haver algum grau de dependência espacial entre as observações.

Tabela 2: I de Moran

	txcrime	ocorrências
I de Moran	0.0118	0.0144
p-value	0.0000	0.0000

Fonte: elaboração própria.

A figura 2, abaixo, também apresenta o conjunto de *hot spots* e *cold spots* áreas -regiões de maior e menor concentração de ocorrências - e o nível de significância, para a medida LISA adotada, o I de Moran Local. Os resultados evidenciam a heterogeneidade de um amplo conjunto de setores com altos índices de criminalidade vizinhos de regiões com baixos índices de ocorrências e vice-versa.

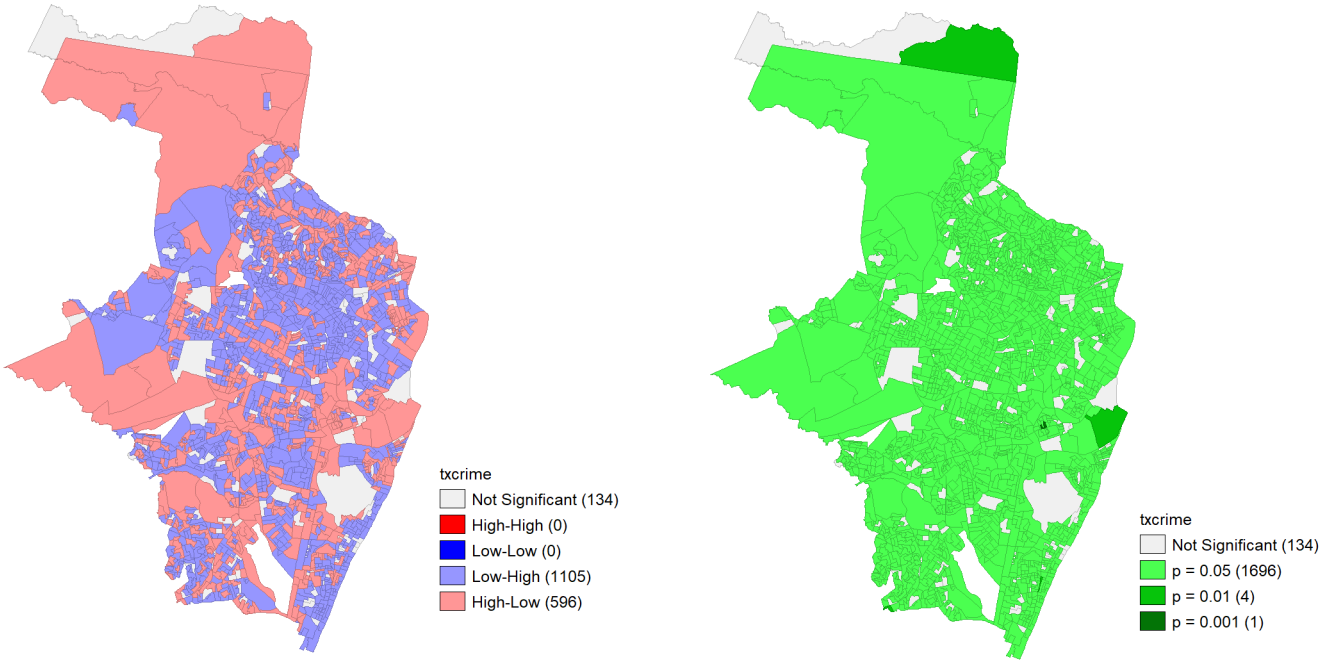

Figura 2: Hot e Cold *spots* áreas e nível de significância para I de Moran Local

A tabela 3, a seguir, apresenta um conjunto de resultados para as estimações dos modelos tradicionais MQO e Logístico nas suas formas convencionais e também para suas expansões considerando as duas formas de dependência espacial consideradas no trabalho (especificações SDM e SLX). Note-se, então, que tratando-se do critério de seleção adotado, MSE de previsão, observa-se que em ambas as formulações, os modelos sem defasagem especial, MQO e Logit, empreendem melhor poder preditivo. Os modelos de variável resposta binária também apreendem melhor a variabilidade dos dados - maior pseudo-R^2.

Tabela 3: Critérios de seleção e performance para os modelos tradicionais MQO e Logit

	MQO	MQO - SLX	ML - SDM	Logit	Logit - SLX	Logit - SDM
Nº de observações	1468	1468	1468	1468	1468	1468
R^2-ajustado	0.169	0.192	-	-	-	-
Pseudo-R^2	0.154	0.167	0.193	0.225	0.268	0.104
MSE	3.277	3.070	2.888	0.191	0.181	0.221
MSE previsão	3.121	3.242	4.149	0.212	0.215	0.313
Estatística-F	6.743***	4.361***	-	-	-	-

Fonte: elaboração própria.

5.3 Modelos penalizados - Variável contínua

A tabela 4 abaixo, apresenta resumidamente, o número de variáveis resultante de cada formulação, o critério de seleção MSE e MSE de previsão, o parâmetro penalizador λ resultado e o indicador de performance dos dados, o pseudo-R^2. Os modelos *ridge 1*, *lasso 1* e *enet 1* correspondem às

estimações sem a consideração das defasagens espaciais, enquanto os modelos *ridge 2*, *lasso 2* e *enet 2* correspondem as estimações considerando as características das vizinhanças - componente especial. Observamos que o critério de *cross-validation* é, em todas as formulações, o que apresenta o menor erro quadrático médio de previsão.O modelo *enet 1* tem performance significativamente superior aos demais - maior poder preditivo relacionado à taxa de ocorrência de CVLI nos setores censitários a um conjunto de 22 variáveis. Note-se, também, que todos os modelos apresentam menor MSE de previsão quando comparados a aqueles da Tabela 3 (modelos tradicionais).

Tabela 4: Critérios de seleção e performance (variável dependente é a taxa de ocorrências de CVLI por 10.000 habitantes no setor censitário)

		nº var.	MSE	MSE previsão	λ	pseudo-R^2
Ridge 1	CV	53	3.5203	1.3448	1.5315	0.1815
	AIC	53	3.3967	1.5173	0.0489	0.1781
	BIC	53	3.3967	1.5173	0.0489	0.1781
Lasso 1	CV	20	3.5463	1.3121	0.07622	0.1734
	AIC	41	3.4022	1.5055	0.0061	0.1773
	BIC	34	3.4179	1.4591	0.0142	0.1734
Enet 1	CV	22	5.4006	0.5538	0.6382	0.1195
	AIC	44	5.4933	1.7304	0.0120	0.1842
	BIC	19	5.3893	0.8354	0.2630	0.1347
Ridge 2	CV	106	3.1865	1.3271	1.3496	0.1843
	AIC	106	3.0377	1.5398	0.0473	0.1856
	BIC	106	3.0377	1.5398	0.0473	0.1856
Lasso 2	CV	30	3.1748	1.3163	0.0508	0.1864
	AIC	93	3.0055	1.6240	0.0004	0.1843
	BIC	40	3.1038	1.3922	0.0200	0.1853
Enet 2	CV	41	3.1305	0.9764	0.1885	0.1885
	AIC	106	2.9964	1.2303	0.0000	0.1856
	BIC	44	3.1035	1.0378	0.0894	0.1885

Variável dependente é a taxa de ocorrências por 10.0000 habitantes. Os modelos Ridge 2, Lasso 2 e Enet 2 referem-se aos modelos com defasagem espacial. Fonte: elaboração própria.

5.4 Modelos penalizados - Variável binária

A tabela 5, abaixo, apresenta de maneira semelhante aos resultados da tabela 4, um conjunto resumido descritivo e de medidas de performance dos modelos com variável resposta binária. Os modelos *ridge 3*, *lasso 3* e *enet 3* representam as formulações sem a consideração da defasagem espacial dos vizinhos, enquanto as formulações *ridge 4*, *lasso 4* e *enet 4* representam o modelo expandido, com consideração das defasagens espaciais. A formulação que apresenta menor MSE de previsão é a *ridge* onde não há seleção de modelo, apenas penalização da magnitude dos coeficientes.

Tabela 5: Critérios de seleção e performance (variável dependente é binária, indicando se houve, durante o período de análise, ocorrência de CVLI no setor)

		nº var.	MSE	MSE previsão	λ	pseudo-R^2	AUC
Ridge 3	CV	53	0.1984	0.2172	0.1043	0.1164	0.7588
	AIC	53	0.1960	0.2157	0.0178	0.1866	0.7647
	BIC	53	0.1960	0.2157	0.0178	0.1866	0.7647

Lasso 3	CV	35	0.3123	0.3807	0.0051	0.1428	0.7583
	AIC	38	0.3123	0.3944	0.0022	0.1413	0.7638
	BIC	8	0.3308	0.0333	0.0333	0.1347	0.7215
Enet 3	CV	35	0.3068	0.3853	0.0090	0.1163	0.7589
	AIC	28	0.3046	0.3944	0.0040	0.1201	0.7637
	BIC	8	0.3160	0.3899	0.0240	0.1129	0.7469
Ridge 4	CV	106	0.3013	0.3013	0.0359	0.1406	0.7702
	AIC	106	0.2997	0.2997	0.1710	0.1417	0.7747
	BIC	106	0.2997	0.2997	0.1710	0.1417	0.7747
Lasso 4	CV	34	0.3111	0.3111	0.0072	0.1410	0.7644
	AIC	41	0.3051	0.3051	0.0072	0.1502	0.7685
	BIC	24	0.3111	0.4311	0.0072	0.1413	0.7577
Enet 4	CV	42	0.3068	1.1512	0.0104	0.1475	0.7654
	AIC	43	0.3046	1.1712	0.0060	0.1510	0.7922
	BIC	25	0.3160	1.1188	0.0200	0.1475	0.7273

Variável dependente é binária, se houve ou não crime no setor. Os modelos Ridge 4, Lasso 4 e Enet 4 referem-se aos modelos com defasagem espacial. Fonte: elaboração própria.

5.5 Consolidação e variáveis

A tabela 6 apresenta os coeficientes estimados das formulações que apresentam, em cada categoria, o menor erro quadrado médio de previsão. Nas colunas 2 e 4 - estimativas MQO e Enet1 - as estimativas referem-se a modelos com variável dependente sendo a taxa de homicídio, já nas colunas 3 e 4 - Logit e Ridge 3 - as estimativas referem-se a especificações com variável dependente binária para a ocorrência de homicídio.

Em relação às evidências obtidas por MQO, aquelas obtidas via *elastic net* apresentam clara redução de número de coeficientes, uma estratégia do método para melhorar a previsão do modelo. Assim, nota-se que algumas das variáveis da abordagem tradicional apresentam-se como muito pouco relevantes para previsão do homicídio. De acordo com as estimativas para os coeficientes da modelo *elastic net*, as variáveis que apreendem características sociais e de infraestrutura domiciliar dos setores censitários são mais relevantes que aquelas associadas à localização dos mesmos. Mais especificamente, por exemplo, os percentuais de crianças/adolescentes e negros e pardos, níveis mais desfavoráveis de escolaridade e de renda são características positivamente associadas à taxa de homicídios dos setores censitários que ajudam na previsão de suas ocorrências, assim como características domiciliares como a condição de ser um setor subnormal (comumente associado à favela), a mais elevada densidade e o percentual de domicílios alugados. Por outro lado e curiosamente, são muito menos relevantes a localização do setor censitário quanto a equipamentos de serviços públicos e a presença ou proximidade de amenidades da cidade.

Tabela 6: Coeficientes dos modelos tradicionais e penalizados conforme critério de seleção - menor erro quadrado médio de previsão

	MQO	Logit	Enet 1	Ridge 3
Variáveis	$\hat{\beta}$	$\hat{\beta}$	$\hat{\beta}$	$\hat{\beta}$
Intercepto	11.906**	-12.224*	-.663	2.118
resp_mulher	-.967	-.303	0	-.051
mulher	-11.744***	1.134	0	-.121
idade0_14	-.485	11.714**	.782	.081
idade15_25	3.542	4.196	0	.684
idade26_35	-	1.672	0	-.107

Tabela 6: Coeficientes dos modelos tradicionais e penalizados (continuação)

	MQO	Logit	Enet 1	Ridge 3
Variáveis	$\hat{\beta}$	$\hat{\beta}$	$\hat{\beta}$	$\hat{\beta}$
idade36_45	-5.361**	-2.150	0	-.072
idade46_55	-2.194	2.633	-.0517	-.687
idade55mais	.296	-	-.001	-.115
alfabetizados_15mais	-4.866	6.413	-.060	-.399
negros_pardos	2.043	-.432	.360	.113
renda_dom_percmed	-.000	.001	-.000	.120
renda_dom_pc05	-1.785	3.997**	.353	-.000
renda_dom_pc025	5.610*	1.741	.736	.348
renda_chefe05	-1.409	-1.469	.556	.519
chefe_semrenda	.715	2.397*	0	-.071
soma_renda10	-.000	.000	-.000	.278
rend_mensal_media	.000	-.001	0	-.000
dom_abastagua	-.430	-.303	0	-.076
dom_saneamento	.006	.231	0	.058
dom_eletricidade	.888	-1.212	0	-.091
dom_coletalixo	-2.385***	-.773	0	-.063
dtiposetor	.072	.224	-.222	.017
dom_10salarios	2.628	.712	0	.144
dom_moradequada	.139	-.060	0	.006
dom_pavimentacao	-.189	-.647*	0	-.122
dom_iluminacao	.384	.876**	0	.131
dom_calcamento	.190	.621	0	.079
dom_esgotoaberto	.017	-.193	0	-.036
dom_arborizacao	-.944	-.872	0	-.105
dom_lixoacumulado	-.141	.3595	.000	.024
morador_pordom	.168	.146	-.162	.006
dom_3oumaisbanheiro	.224	-.788	0	-.067
dom_semhomem	3.498*	.197***	0	.135
dom_alugados	.711	2.501**	.223	.342
casas	.245	.904	0	.144
resp_alfabetizado	-.002	.003	.265	-.000
homem_15maisalfa	-1.165***	2.715***	.339	.381
neg_par_15maisalfa	-1.921	.623	3.015	.103
pessoa_sregistro	68.565	92.991	0	11.311
pessoa_srenda	3.834**	3.174	0	.5063
denskm2	-2.0e-05***	-.000***	.005	.080
areakm2	.405***	.738***	0	.067
dsaude	.293	.164	.058	.051
descola	.312**	.3282*	0	.077
dfronteira	.144	.447	0	-.050
dist_avenidaskm	.042	-.256**	0	.011
dist_cbdkm	.033	.1794**	0	.000
dist_parqpracakm	-.071	-.0242	0	-.022
dist_praiakm	-.072	-.266***	0	-.010
dist_riokm	-.016	-.207**	0	-.006
dist_roadwaykm	.049	-.0761	-.000	.001
dist_subwaykm	.001	.0491	-4.54e-06	-.000
dpolicia	.523	-.001	-4.54e-06	-4.88e-06

Tabela 6: Coeficientes dos modelos tradicionais e penalizados (continuação)

	MQO	Logit	Enet 1	Ridge 3
Variáveis	$\hat{\beta}$	$\hat{\beta}$	$\hat{\beta}$	$\hat{\beta}$
nº de observações	1468	1468	1468	1468
MSE	3.277	0.191	5.4006	0.1960
MSE previsão	3.121	0.212	0.5538	0.2157
Pseudo-R^2	0.154	0.225	0.1195	0.1866

Nível de significância (para os modelos tradicionais MQO e Logit): *** (1%); ** (5%) e * (10%). O modelo Enet 1 refere-se a formulação para variável dependente contínua - taxa de ocorrências. E o modelo Ridge 3 refere-se a formulação para variável dependente binária - *dummy* de ocorrência de CVLI no setor. Fonte: elaboração própria.

Quando se considera a variável dependente binária (indicadora de ocorrência de homicídio no setor censitário), o modelo escolhido (*ridge*) não implica redução de variáveis, mas de valor dos coeficientes. De acordo com os resultados da Tabela 6, há, em geral, semelhança em relação ao sinais das variáveis (embora diferenças também se façam presentes). Note-se, além disto, que, à exceção do sinal do coeficiente obtido para condição de ser um setor subnormal, todos os demais coeficientes estimados são consistentes com aqueles obtidos no modelo da especificação *elastic net* para a variável depende contínua acima discutido. Note-se, porém, que aqui há evidente relevância com respeito à proximidade de equipamentos de serviços públicos de saúde e educação: como se percebe a partir da coluna 4 da Tabela 6, maiores distâncias a tais equipamentos estão positivamente associados à probabilidade de ocorrência de homicídios nos setores censitários da cidade. Por outro lado, ainda que apresente efeitos de magnitudes menores, o distanciamento a amenidades da cidade (praia, rio e parques) tende a reduzir a probabilidade de ocorrência de homicídio.

6 Considerações Finais

Diante, de um lado, do fato de que homicídios ocorrem de forma absolutamente assimétrica no espaço intra-urbano e, do outro, da existência de diferentes teorias que procuram explicar e entender a ocorrência de crimes violentos urbanos a partir de diferentes perspectivas e enfatizando diferentes variáveis, este trabalho utilizou informações de setores censitários, um amplo e inédito conjunto de variáveis e técnicas de *machine learning* para estudar a ocorrência de homicídio na Cidade do Recife. A perspectiva adotada, ao invés de privilegiar abordagens teóricas particulares, permitiu explorar a capacidade de associação de diferentes variáveis (mais de 50 foram tratadas), capturando diferentes aspectos da vida socio-econômicas dos indivíduos das vizinhanças (setores censitários), com a ocorrência de homicídios.

O conjunto de evidências do trabalho apontam para dois resultados principais. Primeiro, observou-se que os modelos que adotam uma perspectiva de aprendizado a partir dos dados *machine learning* foram capazes de gerar melhores(em termos de previsão) que os modelos tradicionais baseados em regressões por MQO ou em que incorporação dependência espacial (SDM e SLX). Segundo e não menos importante, as estimativas obtidas indicam que i) os percentuais de crianças/adolescentes e ii) negros e pardos, iii) os níveis mais desfavoráveis de escolaridade e iv) de renda, v) a mais elevada densidade e vi) o percentual de domicílios alugados são as características espaciais positivamente associadas à taxa de homicídios dos setores censitários que ajudam na previsão de suas ocorrências, Este conjunto de evidência oferece, assim, suporte à Teoria da Desorganização Social e ressalta a importância do *social guardianship* para o entendimento da ocorrência de homicídio na Cidade do Recife.

O trabalho pode e está sendo expandido em duas frentes. Primeiro, é possível adicionar novos elementos associados aos momento da ocorrência (manhã, tarde, noite e madrugada) à análise, uma vez que as ocorrências podem ser extremamente sensíveis às variações do *social guardianship* ao longo

dia. Segundo, embora haja ganhos na consideração do nível de agregação aqui utilizado (setores censitários), para muitas das variáveis e circunstâncias, tal nível pode ser restritivo e levar a resultados muitos dispares entre setores vizinhos e homogêneos, já que a cidade do Recife é caracterizada como macro-segregada do ponto de vista residencial (Oliveira e Silveira Neto, 2015). Neste sentido, a aplicação da mesma estratégia a partir de unidades espaciais maiores (por exemplo bairros) pode beneficiar o poder preditivo de um maior número de variáveis.

Referências

ANDRADE, Rayane Maria de Lima. Configurações de homicídios dolosos em Pernambuco: uma investigação sociológica. Repositório da Universidade Federal de Pernambuco, 2015.

BECKER, Gary S.; MURPHY, Kevin M. A theory of rational addiction. Journal of political Economy, v. 96, n. 4, p. 675-700, 1988.

BECKER, Jacob H. The dynamics of neighborhood structural conditions: the effects of concentrated disadvantage on homicide over time and space. City Community, v. 15, n. 1, p. 64-82, 2016.

CAVALCANTI, Filipe M. Silva; MIRANDA, Filipe P.; NETO, Raul da Mota S. Vizinhança e criminalidade: determinando um rankin de violêcia para os bairros da região metropolitana do Recife. VI Encontro Pernambucano de Economia, 2017.

CAVALCANTI, Filipe M. Silva; MIRANDA, Filipe P.; NETO, Raul da Mota S. Spatial correlation between drugs traffic and violente in Brazil: evidence from urban neighborhoods. 46º Encontro Nacional de Economia, 2018.

CERQUEIRA, Daniel Ricardo de Castro. Causas e consequências do crime no Brasil. Banco Nacional de Desenvolvimento Econômico e Social, 2014.

CERQUEIRA, Daniel; LOBÃO, Waldir. Determinantes da criminalidade: arcabouços teóricos e resultados empíricos. DADOS-Revista de ciências sociais, v. 47, n. 2, 2004.

CERQUEIRA, Daniel; MOURA, Rodrigo Leandro. O efeito das oportunidades no mercado de trabalho sobre as taxas de homicídios no Brasil. Anais do Encontro Associação Nacional dos Centros de Pós-Graduação em Economia. Florianópolis (SC), 2015.

DA SILVA SEABRA, Deborah Maria; NETO, Raul da Mota Silveira; DE MENEZES, Tatiane Almeida. Amenidades urbanas e valor das residências: uma análise empírica para a cidade do Recife. Economia Aplicada, v. 20, n. 1, p. 143-169, 2016.

DA SILVA SEABRA, Deborah Maria; NETO, Raul da Mota Silveira; DE MENEZES, Tatiane Almeida. Amenidades urbanas e valor das residências: uma análise empírica para a cidade do Recife. Economia Aplicada, v. 20, n. 1, p. 143-169, 2016.

FELSON, Richard B. et al. The subculture of violence and delinquency: Individual vs. school context effects. Social Forces, v. 73, n. 1, p. 155-173, 1994.

FLORENCIO, Lutemberg; CRIBARI-NETO, Francisco; OSPINA, Raydonal. Real estate appraisal of land lots using GAMLSS models. arXiv preprint arXiv:1102.2015, 2011.

GIBBONS, Steve. The costs of urban property crime. The Economic Journal, v. 114, n. 499, p. F441-F463, 2004.

GLAESER, Edward L.; SACERDOTE, Bruce; SCHEINKMAN, Jose A. Crime and social interactions. The Quarterly Journal of Economics, v. 111, n. 2, p. 507-548, 1996.

GREENE, William H. The econometric approach to efficiency analysis. The measurement of productive efficiency and productivity growth, v. 1, n. 1, p. 92-250, 2008.

HELWIG, Nathaniel E. Adding bias to reduce variance in psychological results: A tutorial on penalized regression. The Quantitative Methods for Psychology, v. 13, n. 1, p. 1-19, 2017.

LEGGE, Sandra. Youth and violence: Phenomena and international data. New directions for youth development, v. 2008, n. 119, p. 17-24, 2008.

LESAGE, James; PACE, Robert Kelley. Introduction to spatial econometrics. Chapman and Hall/CRC, 2009.

LI, Zhengyi. High Dimensional Model Selection and Validation: A Comparison Study. 2015.

LIGHT, Michael T., HARRIS, Casey T. Race, Space, and Violence: Exploring Spatial Dependence in Structural Covariates of White and Black Violent Crime in US Counties. Journal of Quantitative Criminology. 2012;28:559–586.

MEHTA, Pankaj et al. A high-bias, low-variance introduction to machine learning for physicists. arXiv preprint arXiv:1803.08823, 2018.

MENEZES, Tatiane et al. Spatial correlation between homicide rates and inequality: Evidence from urban neighborhoods. Economics Letters, v. 120, n. 1, p. 97-99, 2013.

MIETHE, Terance D.; REGOECZI, Wendy C.; DRASS, Kriss A. Rethinking homicide: Exploring the structure and process underlying deadly situations. Cambridge University Press, 2004.

OSGOOD, D. Wayne et al. Routine activities and individual deviant behavior. American Sociological Review, p. 635-655, 1996.

QIU, Derek. An applied analysis of high-dimensional logistic regression. 2017.

RODRIGUES, Flávio A. da Cunha; BELMIRO, Célio H. Pereira; Neto, Raul da Mota S. Monocentrismo e estrutura urbana: uma análise empírica para a cidade do Recife. 46º Encontro Nacional de Economia, 2018.

SAMPSON, R. J. e RAUNDENBUSH, S. W. (1999). Systematic social observation of public spaces: a new look at disorder in urban neighbourhoods, American Journal of Sociology, vol. 105(3), pp. 603–51.

SEABRA, Deborah Maria da Silva. Mercado imobiliário e amenidades: evidências para a cidade do Recife. Repositório da Universidade Federal de Pernambuco, 2014.

TIBSHIRANI, Robert. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), p. 267-288, 1996.

VASTRAD, Chanabasayya et al. Performance analysis of regularized linear regression models for oxazolines and oxazoles derivitive descriptor dataset. arXiv preprint arXiv:1312.2789, 2013.

WHEELER, David C. Simultaneous coefficient penalization and model selection in geographically weighted regression: the geographically weighted lasso. Environment and planning A, v. 41, n. 3, p. 722-742, 2009.