Condicionantes da Verticalização dos Centros Urbanos: uma Aplicação de *Machine Learning* para a Cidade do Recife

Andrews Augusto Diniz Barros – PIMES/UFPE
Raul da Mota Silveira Neto – PIMES/UFPE
Célio Henrique Pereira Belmiro – PIMES/UFPE

Resumo: Uma medida importante da relação capital-terra em áreas urbanas é a *Floor-Area-Ratio* (FAR), que fornece a área construída total dividida pelo tamanho do lote. Variações na FAR entre as cidades continuam sendo uma medida pouco estudada na estrutura espacial urbana. Em particular, esse trabalho concentra-se em entender os condicionantes da verticalização no período de 1937 a 2019 para a cidade do Recife. Apesar das regressões lineares serem usadas rotineiramente em estudos de economia urbana, esta estratégia pode não ser útil em situações em que as relações entre as variáveis sejam não-lineares ou complexas. Diante dessa circunstância, métodos não-lineares, como árvores de decisão, podem produzir modelos viáveis, sem necessariamente perder informação em relação a variável de interesse. Neste trabalho, o uso de medidas de importância da variável e gráficos efeito local acumulado gerados por modelos de *Random Forest* são propostos como uma ferramenta prática que pode ser usada para superar esse problema.

Palavras-Chave: Recife. Floor-Area-Ratio. Random Forest.

Abstract: An important measure of the capital-land ratio in urban areas is the Floor-Area-Ratio (FAR), which provides the total constructed area divided by the lot size. Variations in FAR between cities remain an understudied measure in the urban spatial structure. In particular, this work focuses on understanding the constraints of verticalization in the period from 1937 to 2019 for the city of Recife. Although Linear regressions are routinely used in urban economics studies, this strategy may be not useful in situations where the relationships between variables are non-linear or complex. Given this circumstance, nonlinear methods, such as decision trees, can produce viable models, without necessarily losing information in relation to a variable. In this work, the use of variable importance and accumulated local effects graphs generated by Random Forest models are proposed as a practical tool that can be used to overcome this problem.

Keywords: Recife. Floor-Area-Ratio. Random Forests.

Código JEL: C14; R14; R33

1 Introdução

Como apontaram recentemente Lima e Silveira Neto (2020), as cidades brasileiras tem se caracterizado nas últimas décadas por crescimento sobretudo vertical. Entre 2000 e 2010, por exemplo, enquanto que o número de domicílios classificados como unifamiliar cresceu cerca de 2,7%, o número de domicílios classificados como apartamento apresentou uma expansão de 11,9% (IBGE, 2012). Ou seja, os centros urbanos brasileiros são cada vez caracterizados pelo uso mais intensivo do solo e, assim, pela verticalização das construções. A importância desta dinâmica contrasta, porém, com a escassez de trabalhos da literatura econômica sobre os determinantes intra-urbanos da intensidade do uso do solo. Parte desta lacuna pretende ser preenchida com as evidências deste trabalho a partir do estudo do caso da Cidade do Recife.

De acordo com a abordagem econômica, a configuração da estrutura física das cidades reflete as decisões de construtores a partir da tecnologia disponível tendo-se com referência os preços dos fatores (incluindo o do solo) e as preferencias das famílias, ou seja, refletem as decisões de maximização do lucro dos empreendedores (Brueckner (1986), Masahisa Fujita et al. (1989), Lucas e Esteban (2002); Duranton, Henderson e Strange (2015)). Sob tal perspectiva, variações espaciais nas intensidades construtivas e, assim, por exemplo, nos diferentes graus de verticalização do espaço urbano refletem as diferentes condições locais quanto ao acesso ao emprego e a disponibilidade de amenidades. Nesta perspectiva, a tendência de elevação da altura das edificações à medida em que o observador se aproxima dos centros de emprego ou de locais com presença de amenidades observada nas cidades reflete o mais elevado valor da terra nestas localidades, uma vez que tal aproximação significa menores custos de *commuting* ou maior bem estar, levando à utilização de menos espaço e mais capital das edificações (Masahisa Fujita et al. (1989); Brueckner, J.-F. Thisse e Zenou (1999)).

Apesar da literatura empírica a respeito de estimativas de gradientes de preço do espaço e densidade populacional relativos à distância ao centro de emprego derivadas desta abordagem tradicional ser relativamente abundante (ver para um survey McMillen (2006)), trabalhos que considerem especificamente o gradiente da intensidade construtiva são bem mais raros, o que provavelmente está relacionado com a maior dificuldade de obtenção de informações sobre as áreas construídas dos lotes. Recentemente, Barr e Cohen (2014) e Ahlfeldt e McMillen (2018), contudo, forneceram importantes contribuições a respeito da capacidade da abordagem econômica tradicional em explicar as variações da intensidade construtiva, respectivamente, nas cidades de Nova Iorque e Chicago. Como mostraram estes autores, os resultados indicam que parte importante das características físicas dessas cidades podem ser apreendidas a partir da estrutura de incentivo do modelo tradicional e indicam, em particular, a relevância dos preços do espaço e da associada distância às ocupações.

A maioria destes trabalhos sobre a intensidade do solo urbano contudo apresenta uma importante limitação: fazem uso de modelos essencialmente lineares em suas aplicações empíricas num contexto reconhecidamente de espaço heterogêneo. Note-se que mesmo na implementação não-paramétrica da regressão localmente ponderada, a interpretação dos coeficientes dessas estimações continua sendo linear (Brunsdon, Fotheringham e Charlton (1996)), ainda que influenciados por características locais, o que limita a investigação dos determinantes da intensidade do solo urbano uma vez que suas influências podem assumir diferentes padrões de não-linearidades em diferentes pontos no espaço urbano.

O objetivo deste trabalho é, pois, identificar os condicionantes da verticalização na Cidade do Recife a partir da utilização de um conjunto único de informações sobre seus

lotes (que permitem o cálculo da razão área construída/área do lote, a *Floor Area Ratio*) e aplicação da técnica de *random florest*, um caso específico de aplicação de *machine learning* que permite importante flexibilização (e não-linearidades) na forma como as características dos lotes na cidade podem afetar sua intensidade de uso. A partir de seus georreferenciamentos, a estratégia utiliza um grande conjunto de informações sobre características urbanas dos lotes que potencialmente afetam a intensidade do solo urbano, o que inclui a distância ao CBD, as mais importantes amenidades da cidade e diferentes regulações sobre uso do solo na cidade. O trabalho, além de fornecer evidências até aqui ainda não exploradas para caso das cidades brasileiras, também faz um cotejo entre os resultados preditivos obtidos com aqueles gerados através de técnicas convencionais como, por exemplo, estimativas OLS.

Note-se que apesar dos avanços nas aplicações de inteligência artificial em diversas áreas, o tema ainda é pouco tratado dentro da literatura de Economia Urbana e praticamente inexplorado para entender o comportamento da FAR em relação às diversas características que podem influenciar o uso da terra em um contexto urbano tradicional. Na verdade, grande parte dos trabalhos que utilizam ferramentas de inteligência artificial foca em analisar imagens de satélites para tentar prever o comportamento da expansão urbana (Srivastava, Vargas-Munoz e Tuia (2019), Mao et al. (2020) e Chaturvedi e Vries (2021).) e para previsões de preços dos imóveis por meio de *machine learning* (Baldominos et al. (2018) e Winson-Geideman (2018)) e métodos de *deep learning*, como redes neurais artificiais (Nghiep e Al (2001), Selim (2009) e Yacim e Boshoff (2018)), *support vector machine* (Gu, M. Zhu e Jiang (2011) e Wang et al. (2014)) e *random forest* (Hong, Choi e Kim (2020) e Levantesi e Piscopo (2020)).

É oportuno frisar também que, dentre os diversos métodos existentes de *machine learning*, o *random forest* apresenta uma particularidade bastante pertinente para o uso neste trabalho. Primeiramente, é um método com poucos hiper-parâmetros (quantidade de árvores e o tamanho delas), reduzindo problemas de *tuning*. Mais importante, o *random forest* não requer uma especificação detalhada do modelo, sendo bastante perspicaz em casos onde o contexto estudado apresenta uma natureza repleta de heterogeneidades locais, como é o caso do ambiente urbano. Considerando essa última característica, a estratégia de *random forest* tem um potencial enorme para ser um dos métodos mais apropriados para entender como características do ambiente urbano podem influenciar o uso da terra.

Finalmente, ressalte-se que duas importantes particularidades tornam o caso do Recife extremamente bem situado para a análise. Primeiro, sua característica de ser uma cidade bastante antiga e com um intenso processo de urbanização (provavelmente, primeiro centro urbanizado do país) torna possível o uso de informações sobre a FAR para um período bastante longo de tempo, gerando uma capacidade maior de aprendizado do procedimento. Segundo, entre os grandes centros urbanos do país, a cidade apresenta uma quase única variedade de amenidade naturais (com destaque para a presença marcante de praia e rios) e sociais (por exemplo, as Zonas Especiais de Interesse Social antigas) bastante localizadas e criadoras de heterogeneidades urbanas e, assim, sendo fontes potenciais de variações não-lineares na FAR na cidade (Brueckner, J.-F. Thisse e Zenou (1999)).

2 Fundamentação econômica e estratégia empírica

Nesta seção é apresentada a estratégia empírica utilizada na pesquisa. Tal apresentação, contudo, é precedida de uma breve discussão sobre os fundamentos econômicos da verticalização urbana que, ao mesmo tempo, justifica tanto as variáveis escolhidas, como o método específico aplicado no trabalho para apreensão do comportamento da FAR na

Cidade do Recife.

As associações entre o acesso às ocupações e nível de amenidade e o grau de intensidade do uso do solo urbano podem ser percebidas a partir de um modelo simples para o produtor de residências/edificações. Neste sentido, com o objetivo de motivar a apresentação dos resultados da pesquisa, considera-se uma tecnologia com retornos constantes de escala para construção de residências/edificações, $H(K,L) = K^{\alpha}L^{(1-\alpha)}$, onde K e L indicam respectivamente capital e terra urbana, um preço de mercado da unidade de espaço construída, $R_H(r,a)$, e o preço da unidade do solo urbano, $R(r,a)$, onde d representa a distância às ocupações e a captura a disponibilidade local de amenidades, as escolhas do construtor deve maximizar a seguinte função lucro[1]:

$$Max_{K,L}(R_H(r,a)k^{\alpha} - ik - R(r,a))L \tag{1}$$

onde i é o preço unitário do capital e $k = K/L$. As condições para as escolhas ótimas para k e L geram então uma relação entre a intensidade construtiva do solo urbano k e a distância às ocupações e a disponibilidade de amenidades, uma vez que tais condicionantes afetam o preço do espaço construído:

$$k = \left(\frac{\alpha}{i}\right)^{\frac{1}{1-\alpha}} R_H(r,a)^{\frac{1}{1-\alpha}} \tag{2}$$

$$\frac{\delta k}{\delta i} = \left(\frac{\alpha}{i}\right)^{\frac{1}{1-\alpha}} \frac{1}{1-\alpha} R_H(r,a)^{\frac{\alpha}{1-\alpha}} \frac{\delta R_H}{\delta d} < 0 \tag{3}$$

$$\frac{\delta k}{\delta a} = \left(\frac{\alpha}{i}\right)^{\frac{1}{1-\alpha}} \frac{1}{1-\alpha} R_H(r,a)^{\frac{\alpha}{1-\alpha}} \frac{\delta R_H}{\delta d} > 0 \tag{4}$$

Uma vez que $\frac{\delta R_H}{\delta r} < 0$ e $\frac{\delta R_H}{\delta a} > 0$, ou seja, o valor da unidade de espaço construído se reduz à medida que a distância às ocupações (d) se eleva e, por outro lado, se eleva com aumento da disponibilidade local de amenidades (a), temos os resultados tradicionais associadas às escolhas das famílias (Brueckner, J.-F. Thisse e Zenou (1999); M. Fujita e J. Thisse (2013))[2]. Estes resultados indicam que a intensidade construtiva do solo urbano, que pode ser representada pela FAR, deve se reduzir à medida em que o observador se move do centro para periferia e que, para uma dada distância ao CBD, tal intensidade é maior em vizinhanças mais bem dotadas de amenidades.

Há dois importantes pontos a ressaltar para aplicações empíricas desta abordagem tradicional. Primeiro, a abordagem permite um critério claro para a seleção dos determinantes da FAR (ou verticalização): características das localizações que afetem o acesso ao emprego ou que afetem a qualidade da vizinhança. Segundo e por outro lado, tal abordagem não indica claramente como tais influências atuam no espaço urbano, ou seja, não há indicação clara da dinâmica espacial destas influências (taxa espacial de variação, não-linearidades ou descontinuidades).

Neste sentido, o presente trabalho, ancorado na perspectiva acima, primeiro seleciona um conjunto considerável e inédito de variáveis que mensuram acesso ao emprego, diferentes tipos de amenidades (naturais e sociais) e tipos de regulação do solo urbano da Cidade

[1] Ahlfeldt e McMillen (2018) usam uma estrutura semelhante para estudar a relação entre preço do lote e verticalização na cidade de Chicago, construindo um painel de 140 anos com os dados de cerca de 1.700 construções com alta intensidade de uso do solo.

[2] Tais resultados podem ser obtidos facilmente a partir do equilíbrio especial e de uma função utilidade indireta do tipo VRH, $Y - Tr$, a, onde Y é a renda e $T(r)$ é uma função para custo de commuting. Especificamente, $dRH/dr = -T'(r)/s < 0$ e $dRH/da = -Va/VRH > 0$, onde s é a quantidade de espaço demandado.

do Recife, além de características intrínsecas das construções da cidade (tamanho do lote e ano de construção). Tais características são apresentadas a seguir. Em segundo lugar, a pesquisa faz uso de dois métodos distintos para apreensão dos condicionantes da verticalização urbana: Mínimos Quadrados Ordinários e *Random Forest*, um caso específico de técnica de *machine learning*. O modelo MQO é, usualmente, o ponto de partida de abordagens empíricas envolvendo modelos de *machine learning* por duas razões principais. Primeiro, esses modelos tendem a evidenciar, por meio dos resultados da estimação, uma série de problemas nos dados, como multicolinearidade. Segundo, é um objetivo comum das análises que empreendem múltiplas formas funcionais comparar os resultados entre si e com uma abordagem que pode ser considerada padrão, um *benchmark*.

Random Forest, como proposto por Breiman (2001), tem ganhado espaço, dentro das ciências sociais, como destacam Fawagreh, Gaber e Elyan (2014), para estudar uma série de fenômenos sociais (Berk e Bleich (2013), Hong, Choi e Kim (2020), Biau e Scornet (2016) e Levantesi e Piscopo (2020) e Wheeler e Steenbeek (2021)). O método tem sido utilizado como uma alternativa para a obtenção de resultados preditivos mais robustos, apreensão de não linearidades e, mais recentemente, como em Wheeler e Steenbeek (2021), na decomposição de fatores que estão mais ou menos associados as ocorrências do fenômeno de estudo nas regiões de análise.

O modelo se desenvolveu com elementos de duas metodologias preditivas, árvores de decisão e *bagging* (*bootstrap aggregating*). Como argumentam Hastie et al. (2009), árvores de decisão são um método muito popular para diversas atividades envolvendo *machine learning*, sendo invariantes a escala e várias outras transformações das covariáveis, robustas a inclusão de preditores irrelevantes e produzem métodos de fácil interpretabilidade. Apresentam, no entanto, baixo poder preditivo, podendo resultar *overfitting*, isto é, apresentarem baixo viés e alta variância. Ao realizar uma partição recursiva dos dados em grupos cada vez menores, que são mais homogêneas em relação a variável resposta, as árvores de decisão conseguem capturar relação complexas (não-lineares) entre os preditores e resultam em um conjunto de sentenças se-então, como apresentado abaixo:

```
Se Distância a praia < 3 então
|       Se Bairro Histórico == 0 então
|       |       Se Distância a ZEIS > 7 então
|       |       |       Se Distância ao CBD < 5 então FAR = 5
|       |       |       Caso contrário FAR = 4
|       |       Caso contrário FAR = 3
|       Caso contrário FAR = 2
Caso contrário FAR = 1
```

Figura 1: Ilustração de árvore de decisão

A figura acima apresenta, de maneira ilustrativa, uma árvore de decisão para predição do valor da FAR. Assim, para obter o valor predito da FAR de determinado lote, segue-se estrutura de divisão dos dados como apresentada na estrutura se-então, até que se alcance algum dos 5 nós terminais. Neste exemplo simples, observa-se que a praia tem papel fundamental na intensidade do uso solo, os lotes localizados a uma distância maior que 3 km tem um valor predito baixo da FAR. Então, dentro deste primeiro recorte, três variáveis prevem a FAR: a condição de preservação do bairro como histórico, a distância as ZEIS e a distância ao Centro de Emprego. A tabela 1, abaixo, apresenta uma ilustração da FAR predita, com base na árvore de decisão acima e nas características dos lotes. Note que o lote 1, embora relativamente próximo a praia, faz parte de um bairro considerado histórico e tem valor predito da FAR de 2. O lote 2, por outro lado, tem como atributos

um conjunto de características relacionadas a um valor predito mais alto, com FAR predita de 5.

Tabela 1: Ilustração valor predito FAR
com base na árvore de decisão (figura 1)

Variável	Lote 1	Lote 2
Distância a praia	2	2
Bairro Histórico	1	0
Distância a ZEIS	5	8
Distância ao CBD	4	3
FAR (predita)	2	5

Fonte: Elaborada pelos autores. **1.** Os dados apresentados são fictícios e tem o propósito de ilustrar, com base na árvore de decisão apresentada acima (figura 1), o valor predito da FAR dadas as características dos lotes.

Uma alternativa para atenuar a imprecisão identifica nas árvores de decisão é o algoritmo de *bagging (bootstrap aggregating)*, como em Breiman (1996) e Bühlmann e Yu (2002) que consiste na seleção, com reposição, de B amostras aleatórias do conjunto de dados. Para cada amostra, é obtido o ajuste do modelo e então, obtêm-se a média dos B coeficientes - e/ou outras medias. Esse procedimento leva a uma melhor performance do modelo, por que reduz a variância sem aumentar de maneira significativa o viés. Isso significa, aplicando o método às árvores de decisão que, enquanto as predições de uma única árvore são bastante sensíveis a ruídos no conjunto de treino, a média de muitas árvores não é. No entanto, se uma ou mais variáveis são preditores muito correlacionados, eles serão selecionados em muitas das B árvores, e as árvores apresentação alto índice de correlação, comprometendo a robustez dos resultados.

O modelo Random Forest une a capacidade de apreensão das árvores de decisão, com o incremento de performance da abordagem de *bagging* somado a um elemento de aleatoriamente, na seleção do conjunto de covariáveis de cada árvore, que contribuí reduzindo a correlação entre as múltiplas árvores. Como argumenta Breiman (2001) cada árvore cresce de uma amostra bootstrap dos dados originais e usa uma amostra aleatória dos preditores, em cada nó, para construção das árvores, o que resulta em árvores com baixa correlação, reduzindo o *overfitting*. O processo de estimação consistiu em[3]:

1. Divisão dos dados em dois conjuntos distintos, treino e teste, cada um contendo, respectivamente, 70% e 30% dos dados. Este processo é repetido 100 vezes (número de réplicas), com reposição.

2. Obtenção do valor ótimo (*tuning*) do parâmetro *mtry*, que define, em cada novo nó da árvore de decisão, o número de variáveis selecionadas aleatoriamente que estará disponível para sua construção. O valor selecionado, para cada subconjunto de treino, é aquele que minimiza o *out-of-bag error*.

3. Obtem-se o ajuste do modelo para o conjunto de dados de treino, utilizando, para cada amostra, o valor ótimo de *mtry* e os demais parâmetros *default* da biblioteca **ranger**, no **R**.

[3]No caso da formulação MQO, o processo de estimação é o mesmo que o descrito nos tópicos 1, 3 e 4. São criadas 100 réplicas, respeitando a mesma proporções da formulação *Random Forest* de divisão dos dados em treino e teste, para o ajuste do modelo e então obtenção das métricas de performance, apresentadas como o resultado médio das réplicas.

4. Finalmente, os resultados - importância das variáveis, MSE, MSE de previsão, Pseudo-r2 e Pseudo-r2 de previsão, são calculados os e sumarizados para os conjuntos de treino e teste.

Essa acurácia preditiva, contudo, tem um custo na interpretabilidade dos modelos. Quando os preditores de muitas árvores são combinados, e cada predição pode ser resultado de diferentes preditores, em diferentes conjuntos de dados, não é possível destacar uma relação exata entre os preditores e variável resposta. Isto, como argumenta Wheeler e Steenbeek (2021) levanta um dilema comum as estimações envolvendo modelos popularmente conhecidos como caixa-preta - ou *black box models* - acurária ou interpretabilidade?

Identificar por que o modelo prevê determinados lotes da cidade com alta (baixa) intensidade do uso do solo pode ser informativo para a formulação efetiva de políticas públicas orientadas. Um local em que se prevê um padrão mais alto (baixo) de intensidade do uso do solo dada determinada característica X, difere de uma região que prevê um padrão semelhante, dada determinada característica Z, em termos de abordagem. Cabe destacar, contudo, que não se trata de um efeito causal e sim da associação mais (ou menos) relevante, de um conjunto de preditores a ocorrência do fenômeno na região de análise.

Como argumentam Wheeler e Steenbeek (2021) e Molnar (2020), avanços recentes tem permitido reduzir a complexidade dos modelos em métricas que são fáceis de interpretar. Neste trabalho, duas dessas medidas são ilustradas, o nível de importância das variáveis, que permite identificar a redução da acurácia preditiva quando uma variável em particular é permutada aleatoriamente; e os gráficos de efeito local médio, que retratam a relação entre a variação de uma covariável de interesse e a variação no valor predito da variável resposta.

O ranking de importância das variáveis do modelo *Random Forest*, é obtido por meio de uma permutação aleatória de cada covariável - como, por exemplo, a permutação do valor da distância a praia entre os lotes, mantendo todas os demais condicionantes com seus valores originais - para obtenção do erro quadrático médio (MSE), como apresentado por Breiman (2001). As variáveis são rankeadas, então, por ordem de importância para a acurácia preditiva, refletindo o quanto sua não consideração - ou imprecisão - influenciaram no aumento do MSE.

Já os gráficos de efeitos locais acumulados - ou médios - como definem Molnar (2020) e Biecek e Burzykowski (2021), descrevem como as covariáveis influenciam, na média, a predição de um modelo *machine learning*. Apley e J. Zhu (2020) apresentam a formulação teórica para sua construção. Trata-se de uma medida baseada na distribuição condicional e na diferença de preditiva entre intervalos, sendo mais robusta que os gráficos de efeito parcial ou marginal, por não sofrer influencia de outras covariáveis correlacionadas com a de interesse. Como apresenta Wheeler e Steenbeek (2021), efeitos locais acumulados são estimados calculando a mudança na distribuição prevista ao variar ligeiramente X por uma pequena quantidade e, em seguida, calcular a média dessa mudança ao longo de toda a distribuição observada. Por exemplo, calcula-se a distribuição prevista quando o valor de X é substituído por 5, mantendo todas as outras variáveis na amostra em seus valores observados, e chama-se $\hat{p}_{x=5}$, então, aumenta-se o valor para 6, e obtêm-se $\hat{p}_{x=6}$. Finalmente, obtêm-se diferença entre esses dois preditores, $\hat{p}_{x=6} - \hat{p}_{x=5}$ e, em seguida, a média diferenças em todo o conjunto de dados. Essa diferença média é o efeito local quando $X - 6$ e, os gráficos de efeito local acumulado, são a soma cumulativa desses efeitos locais sobre uma grade de valores de X.

Finalmente, vale destacar que a performance geral dos modelos de *machine learning* está relacionada, dentre outros aspectos, com sua capacidade preditiva. Mais especificamente,

com a capacidade dos modelos de realizar boas previsões em novos conjuntos de dados, sendo fundamental a utilização de métricas que permitam aferir de maneira precisa essa performance. Como critério de performance para a comparação das duas diferentes abordagens, duas medidas foram aplicadas neste trabalho, o erro quadrado médio - MSE - e o pseudo-R^2, aplicados nos conjuntos de treino e validação - onde denotam-se, respectivamente, por MSE de previsão e Pseudo-R^2 de previsão.

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 $MSE(\hat{\theta}) = Var(\hat{\theta}) + viés^2(\hat{\theta})$ e calculado como $MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y} - y)^2$. 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, Cribari-Neto e Ospina (2011), que consiste no quadrado da correlação de Spearman entre a variável resposta e os valores preditos. No caso do MSE de previsão, e do Pseudo-R^2 de previsão, as variáveis utilizadas são os dados estimados no conjunto de teste ($\hat{y}_{teste}$) e o valor da FAR neste conjunto y_{teste}.

3 Dados e contexto urbano

O conjunto de dados dos lotes da cidade utilizado no trabalho foi disponibilizado pela Secretaria de Infraestrutura e Serviços Urbanos do Recife (2019) e inclui, entre outras informações, ano de construção e informações sobre as áreas construídas. A partir da base, foi feito um georreferenciamento de todas as observações com fim de ser possível a aplicação do método empírico, além de um minucioso trabalho de tratamento das informações, visto que cada lote urbano pode conter mais de um imóvel, sendo necessário investigar cautelosamente cada lote com o fim de agregar a área construída total, informação relevante para se calcular a FAR do lote e consequentemente mapeamento da intensidade do uso do solo. A base de dados ao fim deste processo é constituída por 97.377 observações de lotes urbanos, incluindo informações de 1874 à 2019, sendo caracterizados por categorias de uso e o ano em que foram construídos.

Os trabalhos que estudam as estruturas espaciais das cidades procuram analisar como a densidade muda em relação ao núcleo da cidade. Para isso, toma-se como base o Marco Zero da cidade do Recife para se definir o centro. Este lugar é considerado, historicamente, como o ponto de partida para fundação da cidade e, como mostrado por Belmiro, Rodrigues e Silveira-Neto (2017), cerca de 59,75% de todos imóveis comerciais da cidade estão localizados em um rádio de 5 km. Ainda, de acordo com Rodrigues, Belmiro e Silveira-Neto (2019), o pico da densidade do emprego é localizado nas vizinhanças do centro histórico, evidenciando o Marco Zero como uma boa referência para definição do centro de negócios e emprego CBD.

A Figura 2, apresentada adiante, além disto, apresenta, além da localização do CBD, duas características distintivas da Cidade do Recife: a presença de praia e rio (Rio Capibaribe). Como mostraram recentemente Seabra et al. (2016) e Lima e Silveira Neto (2020), tais características afetam de forma substantiva o valor dos imóveis da cidade e o tipo de uso destinado aos lotes da cidade (comercial, residencial, multifamiliar). Além da presença de um importante aeroporto, a cidade apresenta ainda diferentes tipos de regulação para uso do seu solo urbano, sendo as mais importantes a presença de sítios de preservação histórica (Recife situa-se entre as mais antigas cidades do país, sendo a mais antiga entre as atuais capitais de estado), sítios com a presença de Zonas Especiais de Interesse Social (ZEIS) e bairros com restrição de verticalização (sob a "lei dos 12 bairros").

A tabela 2, abaixo, apresenta a descrição do conjunto expandido de covariáveis utilizados neste trabalho. Inclui-se na análise os dois principais serviços de infraestrutura da cidade: as estações de metrô e as principais avenidas. Assim como destacado por Masahisa Fujita et al. (1989) e Lima e Silveira Neto (2020), tais serviços são primordiais para o acesso ao local de trabalho (CBD). Finalmente, as amenidades naturais incluem o Rio Capibaribe, a praia e os principais espaços públicos abertos (parques). Como destacado por Brueckner, J.-F. Thisse e Zenou (1999) e Lima e Silveira Neto (2020), as amenidades naturais e históricas desempenham um papel fundamental para a escolha locacional familiar, influenciando o padrão de uso da terra no espaço urbano. Finalmente, inclui-se na análise as Zonas Especiais de Interesse Social (ZEIS), cujo delimitam parâmetros urbanísticos específicos a serem atendidos e que concentram, em geral, a população de baixa renda. Para captar características específicas que referem-se a maior presença de controle urbano na região - bairros - e inerentes ao próprio lote, foram adicionadas variáveis *dummy* que identificam se o lote se encontra em bairro histórico, se o lote faz parte de bairro contemplado na Lei dos 12 bairros[4] e, por fim, se o lote faz parte de uma Zona Especial de Preservação Histórica, definidas no Plano Diretor do município.

Tabela 2: Descrição das variáveis e estatísticas descritivas

Variável	Descrição	Média	Desvio padrão	Mínimo	Máximo
FAR	Floor Area Ratio	1,83	1,49	0,13	5,98
area_lote	Área do lote	6,26	0,90	3,18	12,61
area_lote2	(Área do lote)^2	12,52	1,80	6,36	25,22
year	Ano de construção	1980,65	18,31	1937	2019
dist_cbd	Distância até o Centro de Emprego	5,90	2,24	0,17	13,69
dist_cbd2	(Distância até o Centro de Emprego)^2	39,87	26,69	0,03	187,47
dist_cbd3	(Distância até o Centro de Emprego)^3	293,67	274,93	0,01	2566,82
dist_praia	Distância até o ponto mais próximo da praia	4,02	2,58	0,04	13,78
dist_capibaribe	Distância até o ponto mais próximo do Rio Capibaribe	2,58	2,34	0,00	9,48
dist_parques	Distância até o parque (ou praça) mais próximo	2,75	0,52	0,03	5,96
dist_metro	Distância a estação de metrô mais próxima	2,56	1,77	0,02	12,32
dist_avenidas	Distância até a avenida mais próxima	0,39	0,43	0,00	4,92
dist_aeroporto	Distância até o Aeroporto Internacional do Recife	7,11	3,64	0,05	18,51
dist_ZEIS	Distância até a ZEIS mais próxima	0,42	0,32	0,00	2,83
dhistorico	Dummy indicando se o bairro é histórico	0,06	0,23	0,00	1,00
d12bairros	Dummy indicando se o bairro faz parte da lei dos 12 bairros	0,02	0,14	0,00	1,00
dzeph	Dummy indicando se o lote faz parte de uma ZEPH	0,05	0,22	0,00	1,00

Fonte: A área do lote está em **log do** m^2 - metro quadrado - e as distâncias estão em quilômetros. **2.** No caso das amenidades e equipamentos públicos urbanos que representam, geograficamente, polígonos abrangentes, considera-se a distância até o ponto mais próximo.

Para análise empírica realizada neste trabalho, foram mantidos os lotes que referem-se a edifícios residenciais e de uso misto - apartamentos, edifícios e condomínios - visto que esse tipo de moradia tende a ter um custo associado de demolição relativamente mais alto que lotes com outros tipos de uso, como comércio e casas, por exemplo, reduzindo um possível viés associado aos dados tratarem apenas das construções existentes, desconsiderando possíveis demolições para intensidade mais alta do uso do solo, como destaca Barr e Cohen (2014). Com este recorte, trabalhamos com 9.537 lotes que, dada sua natureza, direcionam a análise em torno da FAR relacionada aos condicionantes da verticalização.

A Figura 2, a seguir, apresenta a distribuição dos lotes considerados e seu nível de FAR no espaço da cidade, juntamente com a localização do CBD, do mar e do Rio Capibaribe. Apesar do tamanho dos lotes dificultar a clara percepção dos diferenciais quanto à FAR,

[4] A lei regula a verticalização dos imóveis, define maior percentual de solo natural para as construções; adapta a altura da edificação à largura da rua; e estabelece padrões de ocupação com formas diferenciadas para proteger a paisagem e a arquitetura dos lugares. Para um entendimento mais aprofundado de como essa legislação impactou na dinâmica da configuração urbana, ver BARROS (2018) e Lacerda et al. (2018)

é clara a maior presença destes nas proximidade do rio e do mar e relativamente menor concentração no entorno do CBD da cidade (uma evidência consistente com a utilização do espaço do CBD apenas para atividade não-residencial, como assinalado também por Lima e Silveira Neto (2020)). Note-se, de forma relevante para esta investigação, que o padrão geral da FAR dos lotes considerados parece longe de se apresentar com um padrão de variação linear em relação a tais características da cidade.

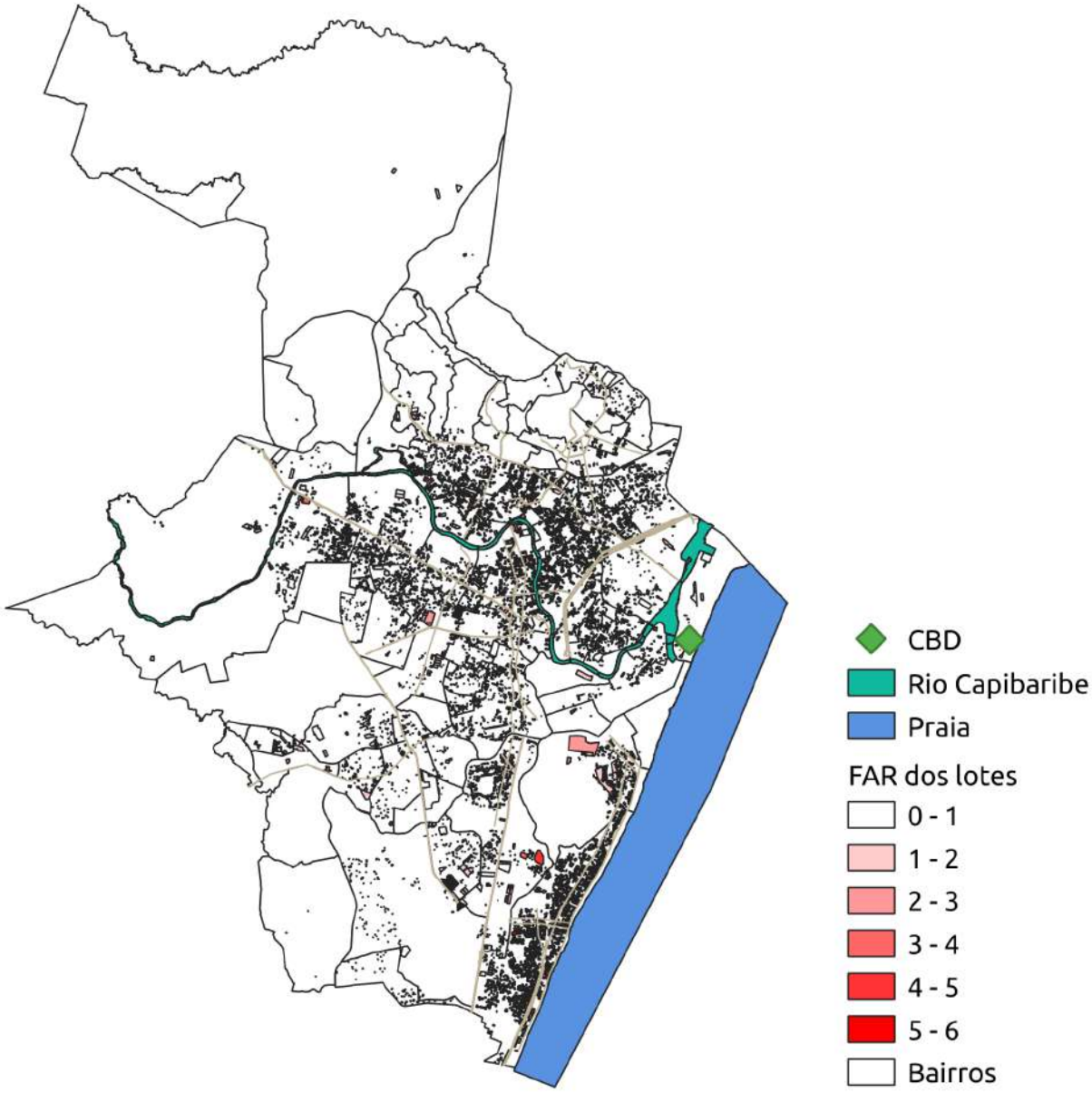
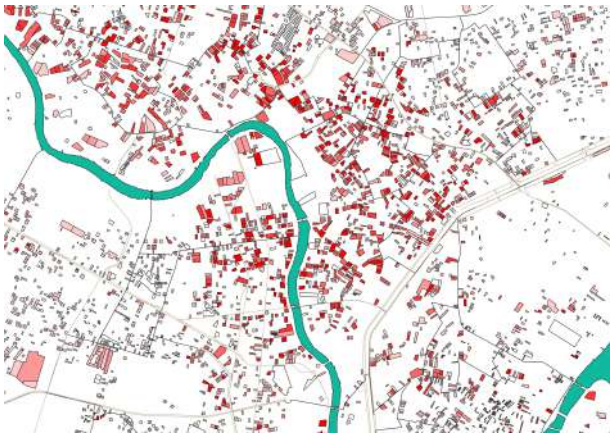


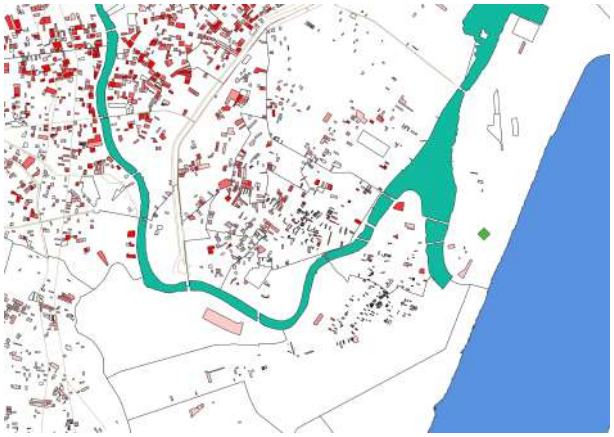
Figura 2: Distribuição da FAR dos lotes dos edifícios residenciais e de uso misto - apartamentos, edifícios e condomínios - e subconjunto de amenidades urbanas - CBD, Praia e Rio Capibaribe.

As Figuras 3(a), 3(b) e 3(c), apresentadas a seguir, ampliam as áreas próximas à praia, rio e CBD, respectivamente, e permitem mais claramente perceber a influência sugestiva da proximidade às amenidades naturais da Cidade do Recife para maior intensidade do uso do solo urbano (maior FAR) e muito menor presença de lotes com eleva FAR nas proximidades do CBD. Note-se, contudo e mais uma vez, que as dinâmicas de variação da FAR em relação a distância a tais características não parece simples ou passível de ser apreendida com especificações empíricas lineares.

(a) Praia de Boa Viagem



(b) Rio Capibaribe



(c) Centro de Emprego (CBD)

Figura 3: Distribuição da FAR dos lotes de edifícios residenciais: figura (a) Recorte ao redor de trecho do Rio Capibaribe; figura (b) Recorte ao redor de trecho da praia de Boa Viagem; e ficura (c) Recorte ao redor do Centro de Emprego (CBD).

4 Resultados

Nesta subseção, são apresentadas evidências básicas a respeito da natureza dos gradientes FAR com respeito ao acesso ao CBD e às amenidades presentes na Cidade do Recife obtidas a partir estimações de um modelo de regressão linear múltipla através do estimador de Mínimos Quadrados Ordinários e Random Forest. É importante deixar claro que diante da estruturação do Random Forest, todos os outros modelos estão estruturados em um formato similar, afim de que haja comparatividade. Todos os métodos expostos nessa seção serão divididos em um conjunto de treino (aprendizagem) e conjunto de teste (validação.). A Figura 2 ilustra os lotes urbanos da cidade do Recife e seus respectivos coeficientes da FAR. É importante ressaltar que diante da limitação da exposição dos dados, não fica claro a ilustração da intensidade do uso do solo na cidade do Recife. Apesar disso, há uma exposição mais detalhada na Figura 3, onde é possível verificar o uso do solo mais intenso em localidades próximas as principais amenidades da cidade, cujo é o Rio Capibaribe e a praia de Boa Viagem. Além disso, pode-se notar um comportamento bastante condizente com os resultados da Tabela 3 no que se diz respeito ao uso vertical do solo em relação ao CBD, onde é notório um uso acentuado próximo ao CBD, mas ao se afastar verifica-se um queda acentuada no uso do solo (Figura 3c).

Os resultados obtidos para o MQO estão listados na Tabela 3 e dizem respeito a três diferentes especificações. A especificação da coluna (1) refere-se a uma estimação simples e básica da influência da distância dos lotes ao Marco Zero. A equação mostra que em média, consistente com o argumento econômico, o gradiente da FAR diminui em aproximadamente 7,8% por quilômetro de distância ao centro. Na especificação da coluna (2), como o objetivo de considerar heterogeneidades espaciais presentes no meio urbano, são adicionados os termos quadrático e cúbico da distância ao Marco Zero (McMillen (2006)). Embora ainda de acordo com a teoria, a significância das estimativas para tais termos indica que a relação entre acessibilidade ao emprego e intensidade construtiva está longe de ser linear e ressalta importância de se considerar heterogeneidades urbanas.

Tabela 3: Resultados do modelo MQO

Variáveis	Modelo 1	Modelo 2	Modelo 3
Intercepto	-187,339	-202,762	-202,611
area_lote	0,304	0,267	0,303
area_lote2	-0,324	-0,831	-1,032
year	24,839	27,072	26,941
dist_cbd	-0,780	-0,513	-0,092
dist_cbd2	0,119	0,117	0,062
dist_cbd3	-0,006	-0,006	-0,003
dist_praia	-	-0,272	-0,286
dist_capibaribe	-	-0,251	-0,229
dist_parques	-	0,280	0,283
dist_metro	-	0,023	-0,034
dist_avenidas	-	-	-0,070
dist_aeroporto	-	-	0,049
dist_zeis	-	-	0,319
dhistorico	-	-	0,259
d12bairros	-	-	0,448
dzeph	-	-	0,092
MSE	0,491	0,401	0,386

Tabela 3 (continuação)			
MSE (previsão)	0,492	0,401	0,386
Pseudo-r2	0,186	0,333	0,349
Pseudo-r2 (previsão)	0,182	0,331	0,347
Nº observações (teste)	6.675	6.675	6.675
Nº observações (treino)	2.862	2.862	2.862

Fonte: Elaborada pelos autores. **1.** A variável dependente é o log da **FAR** dos lotes. **2.** Os resultados apresentados são a média das 100 réplicas de bootstrap.

Por mais que seja inserido variáveis quadráticas e cúbicas no MQO, a estatística referente ao coeficiente apresenta uma relação linear entre essas variáveis e a variável de interesse. Apesar disso, há fortes indícios que o comportamento de uma das variáveis mais importantes do modelo, que é a distância ao CBD, possa ter uma relação não linear com a FAR. Diante disso, a Tabela 4 apresenta o principal resultado até então, que é o Random Forest. A princípio, fica evidente que o MSE do RF apresenta uma queda substancial (em torno de 1/10) comparada com as especificações do MQO. É importante ressaltar que esses resultados se tratam do conjunto de treino, onde tal resultado pode ser justificável pelo problema de overfitting dos dados. Porém, apesar do conjunto de treino apresentar um resultado bastante significativo em relação as estimações do MQO, o resultado para o conjunto de teste cai em relação ao conjunto de treino, ou seja, quando há uma inserção de uma nova observação para a modelagem preditiva (árvores criadas pelo RF) que foi feita no conjunto de treino, ainda há uma redução significativa no MSE em comparação com as demais especificações do MQO, porém em menor magnitude.

Tabela 4: Performance dos modelos

		nº var.	MSE	MSE (previsão)	pseudo-R^2	pseudo-R^2 (previsão)
	Modelo 1	7	0,491	0,492	0,186	0,182
MQO	Modelo 2	11	0,401	0,401	0,333	0,331
	Modelo 3	17	0,386	0,386	0,349	0,347
RF	-	13	0,045	0,224	0,939	0,579

Fonte: Elaborada pelos autores. **1.** A variável dependente é o log da **FAR** dos lotes.**2.** Os resultados apresentados são a média das 100 réplicas de bootstrap. **3.** As variáveis presentes nas distintas formulações MQO são descritas na tabela 3. **4.** Em todas as formulações, o conjunto de treino contém 70% das observações (6.675 lotes) e o conjunto de teste contém 30% das observações (2.862 lotes).

Apesar do Random Forest apresentar uma perfomance melhor em relação ao MQO, pode-se questionar a questão da interpretabilidade do modelo, pois sai de uma estimação simples como a do MQO, onde há uma interpretação simples de cada variável em relação a FAR, e chega em uma estimação puramente computacional, buscando, aparentemente, uma melhor preditividade. Diante disso, a Tabela 5 apresenta o ranking de importância das variáveis, onde é possível verificar o quanto essas variáveis impactam o MSE dada a sua ausência na estimação. A variável mais importante para a predição é a distância a praia de Boa Viagem, importante amenidade urbana do município. Quando esta variável é removida do modelo, o log do erro quadrático médio aumenta em cerca de 29%. Característica intrínseca ao lote, sua área em m^2 e o ano de construção, influenciam em igual magnitude a acurácia preditiva. Dos demais componentes de infraestrutura e amenidades urbanas,

destacam-se o aeroporto, o centro de negócios e o Rio Capibaribe, respectivamente. Menos importantes, contudo, são os condicionantes relacionados ao controle urbano.

Tabela 5: Importância das variáveis Random Forest

Variáveis	(%) aumento no ln(MSE)
dist_praia	0,291
area_lote	0,189
year	0,189
dist_aeroporto	0,169
dist_cbd	0,146
dist_capibaribe	0,134
dist_parques	0,118
dist_metro	0,116
dist_zeis	0,072
dist_avenidas	0,051
dhistorico	0,006
d12bairros	0,005
dzeph	0,003

Fonte: Elaborada pelos autores. **1.** Variável dependente é log da **FAR** dos lotes. **2.** A importância das variáveis foi obtida por meio do método de permutação, como em Breiman (2001).

Embora os resultados referente ao ranking das variáveis sejam bastante intuitivos, eles não proveem uma interpretação similar aos coeficientes do MQO. A Figura 4 apresenta os gráficos de efeito local acumulado das principais variáveis de acordo com o ranking da Tabela 5. Essas figuras podem ser interpretadas como efeitos marginais entre os preditores e a FAR. Esses efeitos são estimados a partir de uma mudança na distribuição predita quando há uma leve variação em determinada variável e, em seguida, calcula-se a média dessa mudança ao longo de toda a distribuição observada (para mais detalhes, ver Wheeler e Steenbeek (2021)).

A Figura 4 apresenta os gráficos de efeito local médio para duas características do lote, área e ano de construção, e para seis condicionantes urbanos, a distância ao centro de emprego; a praia; ao rio Capibaribe; as estações de metrô;, ao aeroporto; e ZEIS. No que se refere ao CBD, o efeito local acumulado possui uma distribuição decrescente, sendo quase que inteiramente constante esse comportamento, porém há uma leve acentuação positiva em torno de 3km ao centro e um efeito mais forte quando a distância é acima de 10km. Isso condiz inteiramente com a teoria econômica tradicional (Masahisa Fujita et al. (1989), M. Fujita e J. Thisse (2013) e McMillen (2006)), onde tal relação teórica e empírica apresentam esse comportamento. No que se observa na relação da praia de Boa Viagem com a FAR, o efeito é ainda mais forte para as regiões próximas a ela, principalmente de regiões que estão localizadas até 2,5km da praia. Esse apontamento mostra que essa amenidade afeta principalmente regiões bem próxima a ela, que é o caso dos bairros de Boa Viagem e Pina. Esse resultado não só condiz com a teoria e estudos empíricos (Brueckner, J.-F. Thisse e Zenou (1999)), mas ainda aponta que o efeito é bem mais concentrado do que comparado com os resultados do efeito local acumulado do CBD.

Uma relação inversa ocorre quando analisa-se o efeito local acumulado do ano de construção do imóvel e sua relação com a FAR. Fica evidente que a função tem um comportamento positivo, apontando que imóveis mais novos tem um efeito maior do que

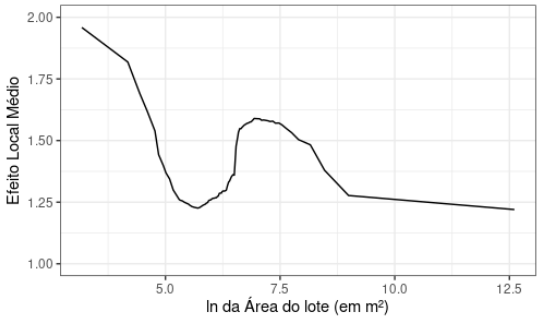


Figura **(a)**

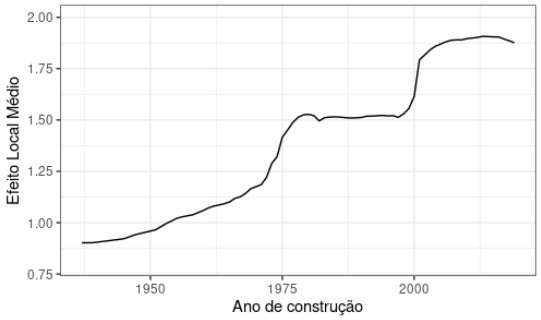


Figura **(b)**

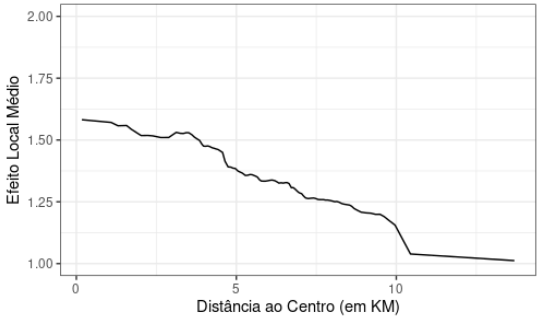


Figura **(c)**

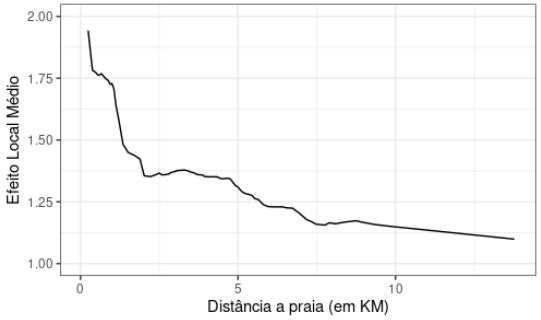


Figura **(d)**

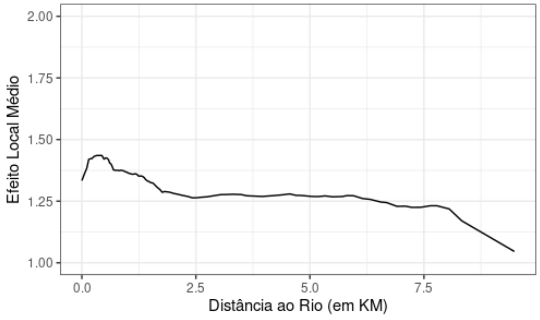


Figura **(e)**

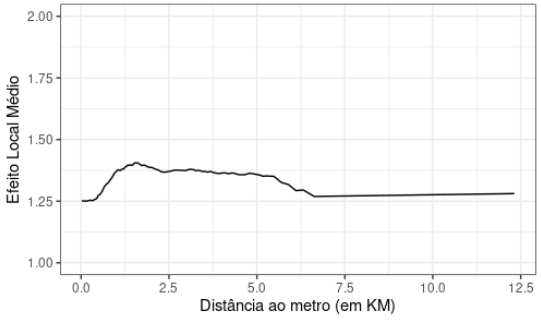


Figura **(f)**

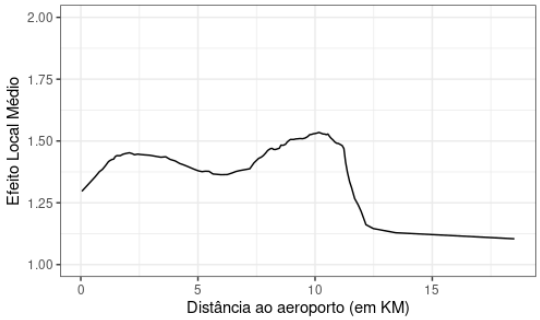


Figura **(g)**

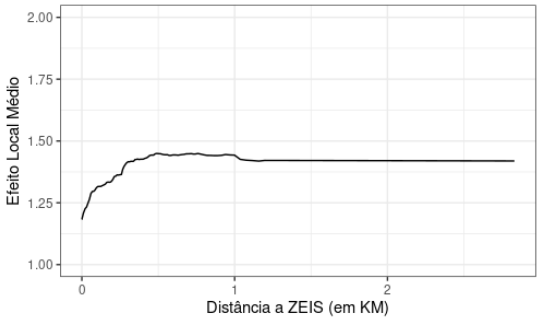


Figura **(h)**

Figura 4: Gráficos de Efeito Local Médio: Figura (a) - **Área do lote**; Figura (b) - **Ano de construção**; Figura (c) - **Distância ao Centro**; Figura (d) - **Distância a praia**; Figura (e) - **Distância ao Rio Capibaribe**; Figura (f) - **Distância ao metrô**; Figura (g) - **Distância ao aeroporto**; Figura (h) - **Distância a ZEIS**.

imóveis antigos. Esse resultado é bastante esperado, já que imóveis mais novos tendem a empregar melhor tecnologia e materiais mais modernos (capital) para aproveitamento do uso do solo. Um ponto interessante não é somente essa relação entre ano de construção do

lotes, mas sim pontos na distribuição em que esse efeito é mais acentuado. Por mais que os imóveis mais novos tenham esse efeito maior, imóveis dos anos de 1975 e 2000 apresentam um efeito bem mais acentuado do que os imóveis anteriores a esses. Há dua potenciais explicações para tais efeitos e, diga-se, bem diferentes. O ano de 1975 é o ano da maior enchente dos Rios Capibaribe e Beberibe na cidade, calamidade que isolou a Cidade do Recife do resto do país por dois dias. Os anos seguintes a 1975 foram de grande procura por apartamentos (que isolavam pessoas de inundações) e eram construídos em lugares específicos da cidade (livres das mesmas). Já a partir de 2000, houve maior crescimento da renda das famílias e expansão do crédito imobiliário, fatores que contribuem para forte procura por novas moradias na cidade, sobre tudo as mais novas. Tal fator pode ter contribuído para a elevação do valor do solo urbano em áreas mais procuradas e, assim, elevado a FAR.

No que se refere à área do lote, temos um comportamento bastante peculiar. Apesar do efeito apresentar uma tendência negativa, ou seja, lotes com área menor apresentam um efeito maior na FAR (são esses os tipos de lotes que apresentam um uso mais intensivo do solo), tem-se uma particularidade principalmente para lotes em torno de 500 a 1.000 m^2, apresentando um efeito positivo. Isso nos diz que esses lotes tem um aproveitamento melhor do uso do solo em relação aos lotes menores entre 150 e 499 m^2, ou seja, uma área maior propicia para esses tipos de construções uma intensidade no uso do solo mais acentuada. Diante disso, algumas teses podem ser extraídas desse conjunto de observações: a) Lotes menores tendem a usar mais intensivamente o solo; b) Lotes entre 500 e 1.000 m^2 apesar de terem um espaço maior para áreas livres, tendem a usar o solo mais intensivamente do que lotes menores entre 150 e 499 m^2; e c) Lotes maiores tendem a não usar o solo tão intensivamente como os demais.

Apesar de não serem variáveis tão importantes como as demais de acordo com a ranking da Tabela 5, a distância às ZEIS e ao Metrô possuem um comportamento bastante similar, apontando que há um efeito positivo em relação a FAR, porém há um comportamento constante ao se afastar desses pontos. O Rio Capibaribe apresenta um comportamento bastante similar ao da Praia de Boa Viagem, evidenciando que localidades próximas ao Rio possuem um efeito maior na FAR do que localidades distantes a ele. A relação entre a distância ao aeroporto e o efeito na FAR é um pouco confuso, porém, é evidente que quanto mais distante, menor o efeito em relação a FAR.

5 Considerações

Tendo-se em vista o acelerado processo de verticalização das cidades brasileiras e o pouco conhecimento acadêmico empírico a respeito dos seus determinantes, esta pesquisa apresentou uma análise empírica dos condicionantes da intensidade do uso urbano para o caso da Cidade do Recife, centro urbano bastante antigo e dotado de claras heterogeneidades espaciais. Duas contribuições estiveram entre os objetivos da investigação: preencher parte desta lacuna de conhecimento existente e explorar a relevância da aplicação de uma técnica de *machine learning* (*random florest*) bem situada para análise onde relações não-lineares se fazem presentes, como tende a ser o caso dos ambientes urbanos.

Obtidos a partir de uma perspectiva de análise de Economia Urbana, os resultados obtidos nesta pesquisa confirmam a relevância dos argumentos dos modelos tradicionais econômicos aplicados às cidades: acessibilidades ao emprego e a serviços de transportes, a amenidades naturais e sociais e diferentes formas de regulação ou restrição do solo urbano exercem efeitos efetivos sobre a intensidade construtiva das edificações da Cidade do Recife e em sentidos previstos pela teoria econômica. Por exemplo e a título ilustrativo e em

ordem de importância, a verticalização deve ser acentuada na cidade quando mais próximo o lote estiver do mar, maior o seu tamanho, mais nova for a edificação, mais distante estiver do aeroporto, mais próximo ao centro do emprego e mais próximo estiver do rio.

Não menos importante, dado seu potencial de apreensão de diferentes influências não-lineares, algo esperado no ambiente urbano, os resultados também indicaram que a estratégia de *random florest* é bastante superior à estratégia tradicional de modelagem linear comumente aplicados em trabalhos sobre a FAR (medida pela correlação entre valores preditos e efetivos da FAR, a precisão obtida via *random florest* é quase duas vezes maior que aquela do modelo de MQO (regressão linear multivariada). Em adição, a estratégia utilizada permitiu i) ordenar por ordem de importância os fatores determinantes da verticalização construtiva do lote, estando entre os cinco mais importantes a distância à praia, sua área, ano de construção e distância ao CBD, e ii) descrever as relações (quase sempre não-lineares) destes efeitos.

Note-se que o conjunto destes resultados representa subsídio absolutamente importante para o planejamento urbano da cidade, em particular para a provisão e planos de expansão da infraestrutura de serviços urbanos da cidade. O trabalho pode e será expandido em ao menos duas direções; estudo da verticalização dos lotes comerciais, considerando os diferentes tipos de usos comerciais e cotejo dos resultados da estrategia aqui empregada com demais técnicas alternativas (mas ainda fundamentalmente lineares quanto aos efeitos das variáveis), como, por exemplo, Regressões Geograficamente Ponderadas.

Referências

Ahlfeldt, Gabriel M e Daniel P McMillen (2018). "Tall buildings and land values: Height and construction cost elasticities in Chicago, 1870–2010". Em: *Review of Economics and Statistics* 100(5), pp. 861–875.

Apley, Daniel W e Jingyu Zhu (2020). "Visualizing the effects of predictor variables in black box supervised learning models". Em: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82(4), pp. 1059–1086.

Baldominos, Alejandro et al. (2018). "Identifying real estate opportunities using machine learning". Em: *Applied sciences* 8(11), p. 2321.

Barr, Jason e Jeffrey P Cohen (2014). "The floor area ratio gradient: New York City, 1890–2009". Em: *Regional Science and Urban Economics* 48, pp. 110–119.

BARROS, Andrews Augusto Diniz (2018). "Densidade urbana e zoneamento: uma avaliação do impacto da Lei dos 12 bairros na cidade do Recife". Tese de mestrado. Universidade Federal de Pernambuco.

Belmiro, Célio, Flávio Rodrigues e Raul Silveira-Neto (2017). "De Centro Exportador a Polo de Serviços – a Cidade do Recif ainda é Monocêntrica? Uma Investigação Empírica sobre sua Atual Configuração". Em: *VI Encontro Pernambuco de Economia*, pp. 1–19.

Berk, Richard A e Justin Bleich (2013). "Statistical procedures for forecasting criminal behavior: A comparative assessment". Em: *Criminology & Pub. Pol'y* 12, p. 513.

Biau, Gérard e Erwan Scornet (2016). "A random forest guided tour". Em: *Test* 25(2), pp. 197–227.

Biecek, Przemyslaw e Tomasz Burzykowski (2021). *Explanatory model analysis: Explore, explain and examine predictive models*. Chapman e Hall/CRC.

Breiman, Leo (1996). "Bagging predictors". Em: *Machine learning* 24(2), pp. 123–140.

Breiman, Leo (2001). "Random forests". Em: *Machine learning* 45(1), pp. 5–32.

Brueckner, Jank (1986). "The Structure of Urban Equilibria: a Unified". Em: *Handbook of Regional and: Urban Economics* 2, p. 821.

Brueckner, Jank, Jacques-Francois Thisse e Yves Zenou (1999). "Why is central Paris rich and downtown Detroit poor?: An amenity-based theory". Em: *European economic review* 43(1), pp. 91–107.

Brunsdon, Chris, A Stewart Fotheringham e Martin E Charlton (1996). "Geographically weighted regression: a method for exploring spatial nonstationarity". Em: *Geographical analysis* 28(4), pp. 281–298.

Bühlmann, Peter e Bin Yu (2002). "Analyzing bagging". Em: *The annals of Statistics* 30(4), pp. 927–961.

Chaturvedi, Vineet e Walter T de Vries (2021). "Machine Learning Algorithms for Urban Land Use Planning: A Review". Em: *Urban Science* 5(3), p. 68.

Duranton, Gilles, Vernon Henderson e William Strange (2015). *Handbook of regional and urban economics*. Elsevier.

Fawagreh, Khaled, Mohamed Medhat Gaber e Eyad Elyan (2014). "Random forests: from early developments to recent advancements". Em: *Systems Science & Control Engineering: An Open Access Journal* 2(1), pp. 602–609.

Florencio, Lutember, Francisco Cribari-Neto e Raydonal Ospina (2011). "Real estate appraisal of land lots using GAMLSS models". Em: *arXiv preprint arXiv:1102.2015*.

Fujita, M. e J. Thisse (2013). *Economics of Agglomeration: Cities, Industrial Location and Globalization*.

Fujita, Masahisa et al. (1989). "Urban economic theory". Em: *Cambridge Books*.

Gu, Jirong, Mingcang Zhu e Liuguangyan Jiang (2011). "Housing price forecasting based on genetic algorithm and support vector machine". Em: *Expert Systems with Applications* 38(4), pp. 3383–3386.

Hastie, Trevor et al. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.

Hong, Jengei, Heeyoul Choi e Woo-sung Kim (2020). "A house price valuation based on the random forest approach: the mass appraisal of residential property in South Korea". Em: *International Journal of Strategic Property Management* 24(3), pp. 140–152.

Lacerda, Norma et al. (2018). *Lei dos 12 bairros: contribuição para o debate sobre a produção do espaço urbano do Recife*. Companhia Editora de Pernambuco (CEPE).

Levantesi, Susanna e Gabriella Piscopo (2020). "The importance of economic variables on London real estate market: A random forest approach". Em: *Risks* 8(4), p. 112.

Lima, Ricardo Carvalho de Andrade e Raul Silveira Neto (2020). "Patterns of urban land use in a developing country: the role of transport infrastructure and natural amenities in Brazil". Em: *Spatial Economic Analysis* 15(4), pp. 441–458.

Lucas, Robert E. e Rossi-Hansberg Esteban (2002). "On the Internal Structure of Cities". Em: *Econometrica* 70(4), pp. 1445–76.

Mao, Wanliu et al. (2020). "Comparison of machine-learning methods for urban land-use mapping in Hangzhou city, China". Em: *Remote Sensing* 12(17), p. 2817.

McMillen, Daniel P (2006). "Testing for monocentricity". Em: *A Companion to Urban Economics. Oxford: Blackwell*, pp. 128–140.

Molnar, Christoph (2020). *Interpretable machine learning*. Lulu. com.

Nghiep, Nguyen e Cripps Al (2001). "Predicting housing value: A comparison of multiple regression analysis and artificial neural networks". Em: *Journal of real estate research* 22(3), pp. 313–336.

Rodrigues, Flávio, Célio Belmiro e Raul Silveira-Neto (2019). "Monocentrismo e Estrutura Urbana: uma Análise Empírica para a Cidade do Recife". Em: *46º Encontro Nacional de Economia* 55, pp. 1–21.

Selim, Hasan (2009). "Determinants of house prices in Turkey: Hedonic regression versus artificial neural network". Em: *Expert systems with Applications* 36(2), pp. 2843–2852.

Srivastava, Shivangi, John E Vargas-Munoz e Devis Tuia (2019). "Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution". Em: *Remote sensing of environment* 228, pp. 129–143.

Wang, Xibin et al. (2014). "Real estate price forecasting based on SVM optimized by PSO". Em: *Optik* 125(3), pp. 1439–1443.

Wheeler, Andrew P e Wouter Steenbeek (2021). "Mapping the risk terrain for crime using machine learning". Em: *Journal of Quantitative Criminology* 37(2), pp. 445–480.

Winson-Geideman, Kimberly (2018). "Sentiments and semantics: a review of the content analysis literature in the era of big data". Em: *Journal of Real Estate Literature* 26(1), pp. 1–12.

Yacim, Joseph Awoamim e Douw Gert Brand Boshoff (2018). "Impact of artificial neural networks training algorithms on accurate prediction of property values". Em: *Journal of Real Estate Research* 40(3), pp. 375–418.