Habitat International 139 (2023) 102896

Contents lists available at ScienceDirect

Habitat International

journal homepage: www.elsevier.com/locate/habitatint

Understanding the land use intensity of residential buildings in Brazil: An ensemble machine learning approach

Célio Belmiro [a], Raul da Mota Silveira Neto [b,*], Andrews Barros [a], Raydonal Ospina [c]

[a] Federal University of Penambuco, Recife, Brazil
[b] Federal University of Pernambuco, Av. Prof. Moraes Rego, 1235 - Cidade Universitária, PE, 50670-901, Recife, Brazil
[c] Federal University of Bahia, Salvador, Brazil

ARTICLE INFO

Keywords:
Floor-area ratio
Machine learning
Random forest
Recife

ABSTRACT

The verticalization of cities impacts the quality of urban life. The empirical investigation of the determinants of the floor-area ratio (FAR) of lots using the traditional econometric approaches, however, has little explanatory power, and research about it using machine learning (ML) is almost nonexistent. This study applies two ensemble machine learning strategies, random forest (RF) and extreme gradient boosting (XGBoost), to investigate the determinants of the FAR of all formally registered multifamily residential lots in the city of Recife, Brazil. Taking into account a collection of key determinants influencing the floor area ratio (FAR), which encompass structural, accessibility, environmental, amenity, and policy variables, the findings reveal that the ensemble random forest approach significantly enhances the explanatory ability of these determinants when compared to conventional strategies like ordinary least squares (OLS) or locally weighted regression (LWR). Although generally in line with traditional urban economic arguments, the evidence also reveals important non-linearities in the effects of the variables on the FAR that are useful for urban planning and public housing policy.

1. Introduction

An important part of the recent urbanization of cities involves the construction ever taller buildings, reflecting a trend toward more intensive use of urban land and the verticalization of buildings both for commercial and residential uses (Ahlfeldt & Barr, 2022b; Ahlfeldt & McMillen, 2018). This increased use of urban land is typically linked to various advantages, including reduced commuting distances, enhanced urban sustainability, the preservation of natural land areas, expanded access to local amenities, and increased agglomeration gains for businesses (Belcher et al., 2019; Danton & Himbert, 2018; Liu et al., 2018; Taecharungroj, 2021).

However, the more intense use of urban land commonly also generates negative externalities associated with noise, pollution, disaster risks and sunlight block, and exhaustion of urban service infrastructure, additionally often favoring social segregation (Bertaud & Brueckner, 2005; Duranton & Puga, 2015; Glaeser et al., 2005; Gyourko & Molloy, 2015; Turner, 2005). In response to higher urban land use intensity, policies are often implemented through land use regulation and height restrictions. However, these measures can result in increased real estate prices and urban sprawl, which in turn create challenges for low-income families in accessing employment opportunities and essential services (Brueckner & Sridhar, 2012; Dantas et al., 2018; Gyourko & Molloy, 2015). An understanding of the factors responsible for the higher urban land use intensity in cities, thus, is fundamental for urban planning and designing appropriate public interventions.

Empirical strategies commonly used to study the intensity of urban land use (i.e., how much is built in each lot's area) are generally based on multivariate linear regressions, splines, and locally weighted regression (Barr & Cohen, 2014; McMillen, 2006) and have limitations related to the number of variables considered, multicollinearity issues, difficulties in capturing urban heterogeneities, and the inability to discover non-linear relationships. It is therefore not surprising that such strategies have weak explanatory power for the variations observed in the intensity of land use within cities. Decision tree-based ensemble machine learning (ML) algorithms, however, appear extremely well suited to overcome most of these limitations. In addition to imposing no functional form and no requirement on data distributions, ML algorithms represent data learning processes that are extremely flexible regarding the number of variables considered and their interactions,

* Corresponding author.
E-mail addresses: celio.henrique@ufpe.br (C. Belmiro), raul.silveirant@ufpe.br (R.M. Silveira Neto), andrews.barros@ufpe.br (A. Barros), raydonal@de.ufpe.br (R. Ospina).

https://doi.org/10.1016/j.habitatint.2023.102896
Received 20 February 2023; Received in revised form 19 July 2023; Accepted 1 August 2023
Available online 18 August 2023

characteristics that are potentially useful to reveal non-linearities associated with the heterogeneities typically present in cities. Difficulties in obtaining appropriate data are likely part of the explanation, but it is somewhat surprising that such strategies have not yet been used to study urban land use intensity.

Here we intend to fill this gap. Using official data of all registered multifamily building lots in the city of Recife, Brazil, and the traditional floor-area ratio (FAR) as a measure of land use intensity, our first objective is to apply two ensemble ML strategies (RF and XGBoost), to understand the variance of the FAR of multifamily building lots in the city space. Our second and no less important objective is to identify the most important factors associated with FAR variation and the specific types of associations present. In a distinctive contribution of the paper, here we consider the key determinants of the FAR and use the unique ability of ML strategies to reveal potential non-linear relationships between the FAR and these determinants. In addition to these general contributions, note that Brazilian cities are well suitable for the present investigation. As revealed by Lima and Silveira Neto (2019), the expansion of multi-family residences in Brazil between 2000 and 2010 exceeded by more than four times that of single-family residences.

2. Literature review

2.1. Fundamental determinants of FAR

According to traditional urban economics, the floor-area ratio (FAR) of urban lots reflects the decisions of profit-maximizing developers on the land/capital mix of buildings based on their relative prices (Ahlfeldt & McMillen, 2014; Brueckner et al., 1987; Duranton et al., 2015; Fujita et al., 1989). Plots with high FAR value indicate the high value of urban land, as profit-maximizing developers prioritize using the cheapest factor. High-value plots in turn reflect the market capitalization of preferences based on two fundamental sets of characteristics of the lots: their degree of accessibility to jobs and services and the set of environmental and amenity factors present. From this perspective, high-FAR lots tend to prevail in urban sites closer to employment centers and with good capacity to access services, along with good environmental characteristics and positive amenities (Brueckner et al., 1999; McMillen, 2006).

More recent investigations have focused on understanding more specifically the height of the tall buildings with the addition of new arguments for valuing the height of the built floor and greater emphasis on the behavior of building costs.[1] Danton and Himbert (2018) and Liu et al. (2018, 2020), for example, emphasized the benefits for firms (signalizing welfare and productivity to clients and workers) and families (better views and less exposure to noise and pollution) associated with the height of buildings. Barr (2010, 2016) and Ahlfeldt and McMillen (2018), on the other hand, highlighted the behavior of building costs (marginal costs increasing with the height) and the role of the evolution of technology on the vertical configuration of cities. More recently, Ahlfeldt and Barr (2022a) proposed a model that incorporates such specific costs and gains associated with the height of the floor and in which developers build taller buildings to take advantage of higher rents.

These new arguments indicate that the value of built land may be affected both by the location of the lot (through the traditional arguments above) and also by its position in terms of height. For residential constructions more specifically, the result is that the value of the lot and its FAR reflect both the degree of accessibility to valued locations and the possibility of enjoying some type of gain in well-being associated with the height (view, security, less noise, etc.). Importantly, these specific considerations of the economics of tall buildings reinforce the relevance of accessibility conditions (since it may facilitate gains from agglomeration) and amenity factors and highlight the role of supply-side and building intrinsic characteristics as determinants of the physical intensity of urban land uses. Particularly, the age of buildings and the area of urban lots are recognized as having an influence on the FAR, since they reflect available technology, architectural trends, and constructive potential (Ahlfeldt & McMillen, 2018; Barr & Cohen, 2014). In complement, Barr (2010) and Combes et al. (2011) argued that the availability of solid bedrock can affect the location of taller buildings.

In addition to these fundamental factors, urban land use regulation has been recognized as an important factor affecting land use intensity (Bertaud & Brueckner, 2005; Brueckner & Sridhar, 2012; Glaeser et al., 2005). Aimed at reducing negative externalities from building density, land use regulation directly (specific FAR restrictions or building height limits) or indirectly (urban growth boundaries, preservation of historical sites, specific regulation for poor areas) affects land consumption and FAR (Geshkov & DeSalvo, 2012; Wassmer, 2006; Zhou et al., 2017).

In comparison with the situation observed for the value of urban proprieties, empirical studies about the determinants of the FAR are not so abundant. Even more scarce are investigations about the direct relationship between land value and land use intensity since these demand past information about urban land values (Ahlfeldt & McMillen, 2018; McMillen, 2006). But the available evidence clearly confirms the importance of accessibility conditions and environmental factors and amenities on the FAR. More specifically, the results indicate that higher FAR and taller buildings tend to be associated with the distance of lots to the central business district (CBD), employment centers, and thoroughfares (Ahlfeldt & Barr, 2022a; Barr, 2010, 2012; Barr & Cohen, 2014; Danton & Himbert, 2018; de Andrade Lima & Neto, 2019; McMillen, 2006). Also, the FAR and the height of buildings are affected by urban environmental characteristics and natural amenities, such as the location of airports, distance to beaches, rivers, and green areas (Barr & Cohen, 2014; Brueckner et al., 1999; Danton & Himbert, 2018; de Andrade Lima & Neto, 2019).

The existing evidence further substantiates that the inherent attributes of buildings, particularly their age and lot area, significantly impact the land use intensity of urban lots (Ahlfeldt & McMillen, 2018; Barr & Cohen, 2014; Danton & Himbert, 2018). Ahlfeldt and Barr (2022a), particularly, have shown that skyscraper development is associated with innovations in construction technology and with economic growth. In addition, urban land use regulation, mainly direct FAR restrictions and height limits, have been shown to have a substantial impact on urban land use intensity of regulated sites (Barr, 2010, 2012, 2013; Bertaud & Brueckner, 2005; Brueckner et al., 2017; Brueckner & Singh, 2020; Brueckner & Sridhar, 2012).

Although important, these empirical works do not explore the possibility of the presence of different types of non-linear relationships between the FAR and its determinants in a multivariate context. For example, while McMillen (2006) and Barr and Cohen (2014) consider, respectively, splines and an LWR strategy to study the relationship between FAR and distance to CBD in a more flexible context, they do so in contexts that are still structurally linear and ignoring other important determinants of the FAR that can also affect it in a non-linear way. The result is that they obtain a rather partial and somewhat imprecise empirical picture of the influences of the FAR determinants. In the present research, we overcome these limitations.

2.2. Understanding the FAR through machine learning

Overall, the available evidence about the determinants of FAR of urban lots is in line with the expectations of urban economics. However, considering the objective of understanding the FAR variation through urban space, two shortcomings are present. First, the works that seek to explain the FAR variation throughout all urban spaces considering the roles of different factors that affect urban land value generally have low

[1] Note that, since the arguments generally assume a unity of land for each building, a tall building means a high-FAR building.

explanatory power, that is, such strategies are not able to apprehend an important part of the FAR variance across lots in the city(for example, McMillen (2006); Barr and Cohen (2014)). These studies usually apply traditional strategies, such as multivariate linear regression, linear splines, or locally weighted regression that barely account for urban heterogeneity present in cities (Atkeson et al., 1997). Second, almost all studies are focused on testing the role of a particular set of determinants of FAR without considering the relative importance of different FAR determinants. In other words, the generated evidence has not allowed for a ranking of the FAR determinants based on their relevance to the variation of FAR within cities. (McMillen, 2006). We use ML techniques to address both limitations in an application involving the city of Recife.

The main reasons for the low performance of traditional methods to explain the variation of FAR in cities are probably associated with urban spatial heterogeneity, in turn, related to both social and natural events. This introduces potential non-linearities in the effect of urban characteristics on FAR and expands the number of determinants of FAR to be considered, tending to bring imprecision, multicollinearity, and omitted variable problems to traditional methods (Hu et al., 2019; McMillen, 2006; Taecharungroj, 2021). These limitations persist also in traditional hedonic price models, which are commonly employed to predict real estate values and comprehend the impact of real estate features on those values. As a result, scholars have increasingly turned to ML methods to predict real estate values and address these drawbacks (Ho et al., 2021; Hong et al., 2020; Kang et al., 2021; McCluskey et al., 2014; Pai & Wang, 2020; Taecharungroj, 2021; Tchuente & Nyawa, 2022; Yilmazer & Kocaman, 2020). However, to the best of our knowledge, no application is available to specifically understand FAR variation within cities and to reveal the relative importance of its determinants.

Typically, ML techniques involve algorithms that, instead of necessarily assuming particular forms of relationships between an outcome variable of interest and its potential determinants, specify mechanisms (or rules) for learning from the information contained in the data (Abidoye & Chan, 2017; McCluskey et al., 2014). The process of learning differs according to the particular ML technique, but the strategies generally involve considering different sets of combinations of determinants and verifying their ability to reflect the behavior of the dependent variable (James et al., 2021). For example, while regularization models (i.e., lasso and net elastic regressions (Zou & Hastie, 2005)) directly adjust the coefficients of the variables in order to improve the models' predictive power (assuming some bias), RF and XBG strategies investigate different combinations of influences without imposing any linear relationship between the combinations of factors and the outcome variable. The ability of ML models to consider a greater number of factors that determine the FAR (overcoming potential difficulties associated with multicollinearity) and their greater flexibility in relation to functional forms, on the one hand, and the local heterogeneities of the urban environment, on the other, suggest that such ML strategies can be very useful to understand the behavior of FAR in urban space (Aydinoglu et al., 2021; Bilgilioğlu & Yılmaz, 2023). Actually, since the present research specifically considers only theory-based essential determinants of FAR, the greatest advantage of applying ML strategies in the current research is associated with the ability of these strategies to capture different types of non-linear relationships between the FAR of the lots and its determinants. A particularly interesting characteristic of these last strategies is that their processes of assembling different decisions about the relevance of the variables provide a way to balance bias and variance, which tends to avoid situations of overfitting (Alfaro-Navarro et al., 2020; Kunapuli, 2022).

3. Empirical strategy

3.1. Urban context and data

Capital of the state of Pernambuco, Recife is the most traditional urban center of Brazil's Northeast region and the core city of the Recife Metropolitan Region. With a population of about 1.7 million people and an area of 218,843 km2, the city is the ninth most populous and has the fourth-highest demographic density among Brazilian capitals (https://cidades.ibge.gov.br/brasil/pe/recife/panorama). Besides a port that has been important for commerce since the colonial period and a traditional CBD, Recife is geographically marked by the presence of the Capibaribe River.

Using official data on registered lots in the city in 2019 obtained from the city hall, Fig. 1 presents the location of the set of the 9547 multifamily residential buildings in our database (boundaries define districts), together with the locations of the city's CBD[2] and airport and two distinctive natural characteristics, the beach and the Capibaribe River. Note that the buildings are not very far from the CBD and tend to be close to these natural areas, suggesting their roles as amenities. Fig. 2 presents the location of buildings in groups of the 25% highest-FAR (map in Fig. 2a) and the 25% lowest-FAR buildings (map in Fig. 2b) buildings. These indicate that in general, the highest-FAR buildings are located closer to the beach area and the river course than the lowest ones, suggesting their potential roles as natural amenities. The highest-FAR buildings also tend to be closer to the CBD but not so close to the airport.

Given the unavailability of information on prices of urban lots and motivated mainly by the traditional urban economic empirical literature discussed in subsection 2.1, we consider a total of 13 factors potentially associated with the FAR of lots in the city, distributed in four dimensions of influences.[3] structural or lot intrinsic factors (area and age of the building) (see Barr & Cohen, 2014); accessibility (distances to the CBD, the nearest thoroughfare, and the nearest subway station) (Ahlfeldt & Barr, 2022b; McMillen, 2006); environmental and amenity factors

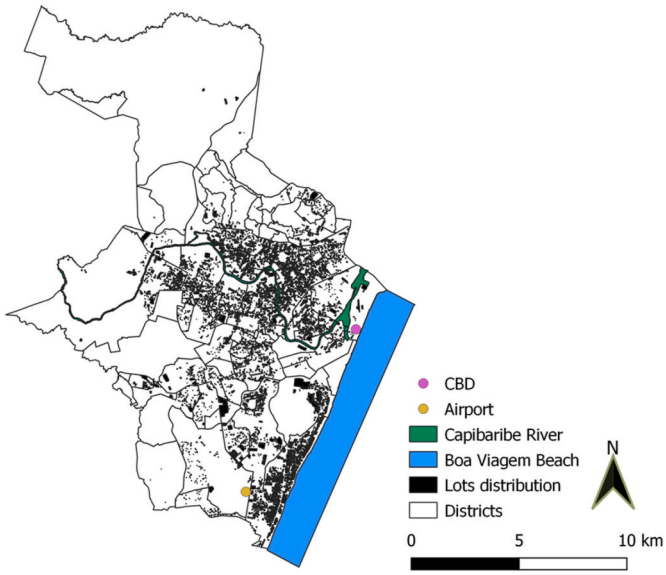


Fig. 1. Spatial distribution of residential buildings in the City of Recife.

[2] Similar to Seabra et al. (2016), the exact location of the CBD of Recife corresponds to the ground zero of the city. Located in the old district of Recife and near the harbor, about 60% of all commercial buildings are located up to a radius of 5 km from it (Belmiro et al., 2018).

[3] Note that the number and the nature of FAR determinants allow an informative and useful comparison of the results with those from traditional multivariate regression approaches, besides being theoretically well established and attenuating more obvious problems of reverse causality. In Fig. 9 and Table 3 of Appendix A, we present a matrix of Spearman correlations between the variables and no strong association is identified.

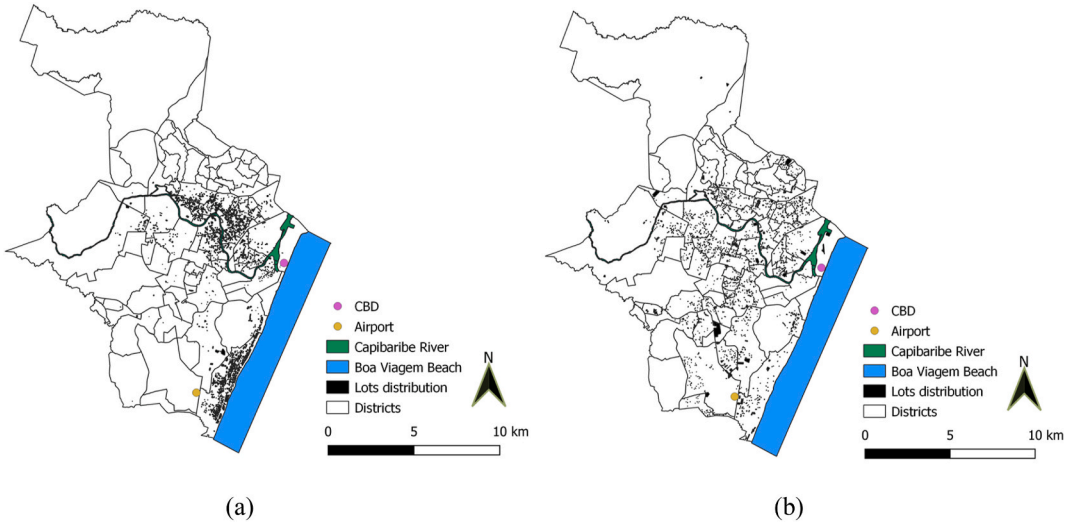


Fig. 2. Spatial distribution of residential buildings in the City of Recife - highest (a) and lowest (b) FAR distribution quartiles.

(distances to the beach, the Capibaribe River, the nearest park, and the airport) (Barr & Cohen, 2014; Danton & Himbert, 2018); and public policy urban interventions (distance to the nearest Special Zone of Social Interest - ZEIS, location in a historic preservation district, location in an area of building height restriction - a contiguous area of 12 districts, and location in a Special Zone for the Preservation of Historical-Cultural Heritage - ZEPH) (Bertaud & Brueckner, 2005; Brueckner & Singh, 2020).[4] We obtained this entire set of information on the characteristics of the lots and the location of social and natural events from official records of the city hall (https://esigportal2.recife.pe.gov.br/portal/apps/webappviewer). The information was then georeferenced, and when applicable, Euclidean distances were calculated. Table 1 presents the descriptions of these variables together with descriptive statistics.

According to the numbers of Tables 1 and in comparison with buildings in the lowest-FAR quartile, highest- FAR quartile ones are newer (the average ages are, respectively, 87 and 34 years), closer to the CBD, subway stations, and thoroughfares, and as mentioned, closer to the beach (but not to the river or parks). The numbers of Table 1 also suggest that public urban policies matter: highest-FAR quartile buildings are also located more distant from a ZEIS and less present in a ZEPH than the lowest-FAR ones.

3.2. Ensemble machine learning: Random forest and extreme gradient boosting

In this investigation, two ensemble ML strategies are considered, ensemble random forest (RF) and ensemble extreme gradient boosting (XGBoost). Their flexibility in capturing non-linearities and the possibility of identifying and ranking relevant determinants of FAR initially justify the choice. In the literature, studies have shown that these methods are competitive both regarding regression and classification (Fan et al., 2018; Fernández-Delgado et al., 2014). Furthermore, compared to traditional linear multivariate models, these two strategies have presented good performance in predicting property prices in urban environments (Iban, 2022; Taecharungroj, 2021).

Random Forest is an extension of the decision tree (DT) model in which estimates are derived from uncorrelated tree set predictions. Decision trees are made up of numerous resamples with replacement made by bootstrap (Breiman, 2001; Friedman, 2001a). Some of its

Table 1
Descriptive statistics - All buildings (All) and the buildings of the two extreme quartiles of the FAR distribution (the top 25% and the bottom 25%).

Variable	Description	All buildings	The top 25%	The bottom 25%
FAR	Floor Area Ratio	1.83 (1.49)	4.10 (1.11)	0.56 (0.16)
lot area	Lot area	6.26 (0.90)	6.75 (0.88)	6.08 (0.85)
Year	Construction year	1981 (18.31)	1988 (17.39)	1977 (18.65)
dist cbd	Distance to the Central Business District	5.90 (2.24)	5.47 (2.26)	6.29 (2.14)
dist beach	Distance to Boa Viagem Beach	4.02 (2.58)	2.79 (2.36)	4.91 (2.32)
dist river	Distance to Capibaribe River	2.58 (2.34)	2.78 (2.68)	2.47 (2.18)
dist park	Distance to the nearest park	2.75 (0.52)	3.01 (0.45)	2.59 (0.52)
dist subway	Distance to the nearest subway station	2.56 (1.77)	2.43 (1.46)	2.90 (1.97)
dist avenue	Distance to the nearest avenue	0.39 (0.45)	0.32 (0.40)	0.45 (0.48)
dist airport	Distance to the airport	7.11 (3.64)	6.55 (3.74)	7.79 (3.56)
dist ZEIS	Distance to nearest ZEIS	0.42 (0.32)	0.51 (0.29)	0.30 (0.30)
dhistoric	Dummy indicating whether the lot is in a historic preservation zone	0.05 (0.23)	0.08 (0.27)	0.01 (0.13)
d12districts	Dummy indicating whether the lot is part of the law of 12 districts	0.02 (0.14)	0.07 (0.26)	0.00 (0.03)
Dzeph	Dummy indicating whether the lot is In a ZEPH	0.05 (0.22)	0.07 (0.25)	0.02 (0.14)

Source: the set of information on the characteristics of the lots and the location of social and natural events from official records of the city hall. Note that: a) Except in the case of dummy variables (the bottom three lines), values represent means; for dummy variables they represent shares. Standard deviations are in parentheses. b) The top 25% and the bottom 75% refer to the distribution of the FAR values of the buildings. c) The lot area is measured in natural log of square meters and the distances in kilometers. d.) In the case of amenities and urban public equipment that represent geographically wide polygons, the distance to the nearest point is considered.

[4] Lots located in a ZEIS cannot be merged in order to rebuild a taller building, and buildings located in historic preservation districts or a ZEPH have to maintain their original facades.

advantages are the robustness to outliers, in addition to presenting low bias and being able to capture complex data interactions. The RF algorithm has the following steps: i) draw bootstrap samples for the number of trees from the original data, ii) grow unpruned regression trees for each bootstrap sample and at each node randomly sample the number of predictors and choose the best division among those variables, and iii) predict the output of test data by the averages of the predictions from all the trees.

Extreme gradient boosting (XGBoost) is an ensemble method that uses the decision tree concept associated with the gradient boosting structure (Breiman, 1996; Friedman, 2001b). XGBoost was initially introduced to improve the training time of Gradient Boosting Machine models and it combines the estimates of a set of decision trees (weaker learners) to predict the output of a target variable (Chen & Guestrin, 2016). In the learning process applied to each sample, XGBoost minimizes a regularized loss function, and new decision trees are added iteratively, one by one, in order to correct the prediction of the previous trees in the model. Particularly, the XGBoost algorithm optimizes the boosting process through an efficient algorithm to calculate the regularized gradients and Hessians of the Taylor series approximation of the loss function. This allows for faster training and improved model convergence to obtain the final result. After all resampling and associated models, this XGBoosting's result corresponds to a weighted combination of outcomes where the weights are adjusted according to the prediction performance of the models.

The XGBoost algorithm has several hyperparameters, such as the number of trees, the depth of each tree, and the number of data points used to train the model. To control the number of data points used for this purpose, it is possible to combine a XGBoost and k-fold cross-validation (Arlot & Celisse, 2010). The Random Forest and XGBoost models were applied, respectively, using the ranger and xgboost packages of the R statistical software, respectively. Given the good performances on a variety of datasets (Liaw et al., 2002; Chen & Guestrin, 2016) and for clarity in the procedures, similarly to (Li et al., 2021; Lin et al., 2022; Wheeler & Steenbeek, 2021), we used the default parameters of both packages.[5] For the RF strategy, we used 1000 trees; as for the XGBoost model, we used a total of 100 of boosting rounds. Note that, in all strategies applied in the research, the results are the average of 100 bootstrap replicates and the training and test samples contain, respectively, 70% and 30% of the total of observations, which amount, correspondingly, to 6675 and 2862 lots.

Despite the characteristics discussed in relation to the gains from the application of the RF and XGBoost algorithms compared to traditional regression techniques (OLS and LWR), we compared the individual performance of the models in the predictions in relation to the use of the batch.[6] To do this, we applied five performance indicators: the mean squared error (MSE), the mean absolute error (MAE), the root mean squared error (RMSE), the coefficient of determination (R^2), and a pseudoR^2. These indicators are described in equations (1)–(5) below:

$$\mathrm{MSE} = \sum \left(y_{i,o} - y_{i,p}\right)^2 \tag{1}$$

$$\mathrm{MAE} = \frac{1}{n}\sum \left|y_{i,o} - y_{i,p}\right| \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum \left(y_{i,o} - y_{i,p}\right)^2} \tag{3}$$

$$R^2 = 1 - \frac{\sum \left(y_{i,o} - y_{i,p}\right)^2}{\sum \left(y_{i,o} - \overline{y_0}\right)^2} \tag{4}$$

$$\mathrm{Pseudo} - \mathrm{R}^2 = \sum cor\left(y_{i,o}, y_{i,p}\right)^2 \tag{5}$$

A similar set of indicators is used, for example, by Hu et al. (2019) and Taecharungroj (2021). Here, $y_{i,o}$ represents the observed FAR value of lot i, $y_{i,p}$ represents the predicted FAR value for the same unit, and $\overline{y_o}$ represents the average value of FAR. The Pseudo-R^2 follows the definition presented in Florencio et al. (2011), who define the metric as a result of the Pearson correlation, squared, between the observed and predicted values. Finally, n is the total number of lots.

3.3. The relevance of variables and their relationships with the FAR of buildings

In a RF model, a set of tools used to measure the importance of variables to describe how much the predictive performance of a model depends on the information of each covariate considered is known as feature importance analysis (Fisher et al., 2019; Wei et al., 2015). In this work, we determined the importance of the variables through the random permutation of each covariate, keeping the others with their original values and then evaluating the results of the mean squared error (MSE) variation, in a similar way to the results presented, for example, by Breiman (2001), Taecharungroj (2021), and Liaw et al. (2002). The variables are ranked in order of importance for the predictive accuracy of the model, reflecting how much their non-consideration influences the consequent increase in the MSE.[7]

While the importance of variables provides a ranking of the importance of predictors for the model, this does not represent an interpretation relative to a regression coefficient (Wheeler & Steenbeek, 2021). In the absence of this analysis, the predictive effect signal rests on theoretical conjectures. To deal with this issue and shed light on the behavior of the relationships between the FAR and the set of predictors considered, we used the accumulated local effect plots (ALE plots). These plots allow visualizing marginal effects by plotting covariates against the predicted outcomes if the features are highly correlated (Apley & Zhu, 2016). This construction is part of a set of techniques called global model-agnostic methods, which also include, among other performance metrics, partial effects plots (PDP) and marginal effects plots, in a way that allows both the data scientist and the end user to interpret the ML results (Apley & Zhu, 2016; Gromping, 2020; Molnar, 2020; Murdoch et al., 2019).

More specifically, based on the conditional distribution of features, ALE calculates predictive differences in small blocks of variation of the covariate of interest, keeping all other covariates at their original values. As an example, suppose we want to know the ALE of lots located 5 km from Boa Viagem beach. In this case, we select all lots that are about 5 km from the beach and obtain the difference from the forecast if they were 6 km away minus the forecast if they were 4 km away. This would return the pure effect of the studied distance of interest, not mixed with the effect of the correlated covariates. These values are then accumulated and centered, resulting in a graph of accumulated local effects

[5] Specifically, for the RF, we used the values 3 and 5, respectively, for the number of variables to randomly sample the candidates at each division and the minimum number of samples within the terminal nodes; for the XGBoost model, we considered a maximum tree depth for base learners of 6, a learning rate of 0.3, a control regularization parameter (to prevent overfitting) of 0.0, and a value of 1 for the minimum number of instances required in a child node for the number of samples and for the number of features supplied to a tree.

[6] Note that Random Forest and XGBoost algorithms make decisions based on relative feature values and thresholds, rather than the absolute values. Therefore, the scale of the features does not affect their performance. On the other hand, OLS (Ordinary Least Squares) and GWR (Geographically Weighted Regression) strategies assume that the features have a similar scale, and large variations in feature scales can lead to important coefficient variations and inaccurate predictions. In such cases, data scaling/normalization is often recommended. In this work and for the purpose of fair comparison the data were not scaled/normalized.

[7] Other adjustment assessment metrics, such as R^2, for example, could also be used, but we chose to follow the works that are referenced.

(Apley & Zhu, 2016; Molnar, 2020; Wheeler & Steenbeek, 2021). Different from other alternatives, such as partial dependence and marginal whose applications fail in the presence of correlation between the predictors, cumulative local effects are unbiased and can be applied when the covariates of interest are correlated (Molnar, 2020). In addition, ALE plots are easy to interpret and, conditional on a given value of the covariate of interest, the relative effect of changing the feature on the prediction can immediately be read from the plot.

4. Results

4.1. RF and XGboost performances

In order to highlight the relevance of our ML strategies (random forest and extreme gradient boosting) in understanding the FAR variance of buildings, we initially compared their performances with those of the traditional OLS regression and of a local weighted regression (LWR). The results of the performance of these models are presented in Table 2 both for the training and testing samples using traditional indicators.[8] The values of the performance indicators for the testing samples show the clear superiority of the two ensemble ML approaches over the more traditional strategies in understanding FAR variations across multifamily residential buildings in Recife.

More specifically, considering the MSE and MAE measures in the test sample, for example, the evidence indicates that the ensemble RF has values of these indicators that are, respectively, 57.4% and 70.6% of the corresponding values obtained by OLS. As for the R^2 values, the ensemble RF presents a value (0.625) about 80% higher than the one observed for the traditional OLS estimator (0.337). Due to the capture of some urban heterogeneities by the LWR strategy, the gains of the two ML approaches over this strategy are lower but still clearly important. For example, the R^2 using ensemble RF is 25.7% higher and its MSE is just 75% of that obtained by the LWR. In Appendix B, using the testing samples, we present scatterplots with actual and predicted FAR values for the OLS and RF strategies that illustrate the best fit of the ML strategy.

Despite presenting more similar values for the performance measures, the numbers of Table 2 also indicate that the parallel ensemble strategy of RF performs better than the sequential ensemble of XGBoost.

Table 2
FAR of residential buildings – Models' performances.

OLS		LWR	XGBoost	RF
Training sample				
MSE	0.392	0.276	0.059	0.046
MAE	0.496	0.405	0.179	0.157
RMSE	0.626	0.525	0.244	0.214
R^2	0.348	0.540	0.901	0.924
Pseudo-R^2	0.342	0.492	0.887	0.940
Test sample				
MSE	0.392	0.300	0.244	0.225
MAE	0.496	0.417	0.364	0.350
RMSE	0.626	0.547	0.494	0.474
R^2	0.347	0.499	0.593	0.625
Pseudo-R^2	0.340	0.465	0.548	0.579

Notes: a.) The dependent variable is the log of FAR lots and the 13 explanatory variables are those presented in Table 1 b.) In all strategies, the results are the average of 100 bootstrap replicates and the training and test samples contain, respectively, 70% (6675 lots) and 30% (2862 lots) of the observations.

[8] In the LWR approach, weights are based on the locations of buildings relative to the CBD through a Gaussian kernel function using an optimal bandwidth (it minimizes the mean integrated squared error). Note that to properly compare strategies, in all cases the values of the indicators represent the average of 100 bootstrap replicates.

The value of 0.625 for the R^2 in the case of RF indicates that 62.5% of the variance of the FAR (in log.) is explained by the set of variables. This value of the R^2 is also higher than the counterparts obtained, for example, by Barr and Cohen (2014) using 16 explanatory variables in a traditional linear OLS specification to explaining the (log. of) FAR in New York City (about 0.44). Actually, our value is similar to those obtained recently by Taecharungroj (2021) using RF and XGBoost strategies to explain condominium prices in Bangkok, and a little lower than the ones generated by Iban (2022) by applying these strategies to explain real estate values in Yenisehir district, Turkey.

Given its superior performance, we use the ensemble RF strategy to discuss the relevance of the variables and measure their effects on the FAR.

4.2. The relevance of the lot characteristics

Figs. 3 and 4 present evidence about, respectively, the absolute and relative importance of the 13 variables for FAR determination. The importance is measured in terms of MSE variation and the variables are highlighted by color according to their nature (accessibility, environmental and amenities, structural, and public policy). In general, the set of evidence indicates the importance of variables associated with the environment and amenities, accessibility, and structural lots' characteristics, and a less relevant role of public policy indicators in lots' FAR behavior.

Specifically, the evidence indicated that the most important variable for FAR determination is the distance to the beach, whose permutation brings an increase in the MSE of almost 30%. As shown in Fig. 4, no other factor presents more than 65.1% of its importance. The other environmental and amenity variables (distance to airport, river, and park) are all also relevant, but they all present less than 60% of the relevance of the distance to the beach. Notice that structural or intrinsic lots' variables (area and year) are also relevant for the FAR behavior in Recife: both present increases in MSE of around 20%, representing about 65% of the variation observed for the distance to the beach.

As for accessibility variables, the most important variable is the distance to the CBD, with about half of the importance of distance to the beach. On the one hand, this evidence is consistent with the idea that the use of cars makes the proximity of employment relatively less relevant, with families prioritizing other aspects in choosing where to live. On the other hand, the continuing relevance of distance to the CBD conforms well with the initial structure of the city and suggests its resilience.

4.3. The relationship between FAR and lot characteristics

The set of Figs. 5–8 presents ALE plots for variables associated with buildings' intrinsic characteristics, accessibility, environmental and amenities, and urban public policies, respectively. ALE plots represent accumulated variations in the predictive FAR values at local values of the variables and can capture different kinds of relationships between FAR and its determinants.

For the two building intrinsic or structural characteristics, Fig. 5 reveals clear non-linear relationships between the predicted FAR variation and the values of these factors. While the predicted FAR variations tend to increase non-linearly with the local value of buildings' age (confirming expectations), the variation of the predicted FAR with the buildings' lot area increases or decreases depending on the local value. Interestingly, Fig. 5a indicates that the FAR variation decreases at lower and higher values of buildings' lot area and increases at intermediary values.

As for accessibility variables, in line with theoretical expectations, Fig. 6(a)–6(c) indicate that, in general, better access to jobs and thoroughfares is associated with high values of predicted FAR variations. Consistent with the initial monocentric character of the city, note also that among the three accessibility factors, the strongest FAR variation effects arise from the variable measuring the distance to the CBD, which

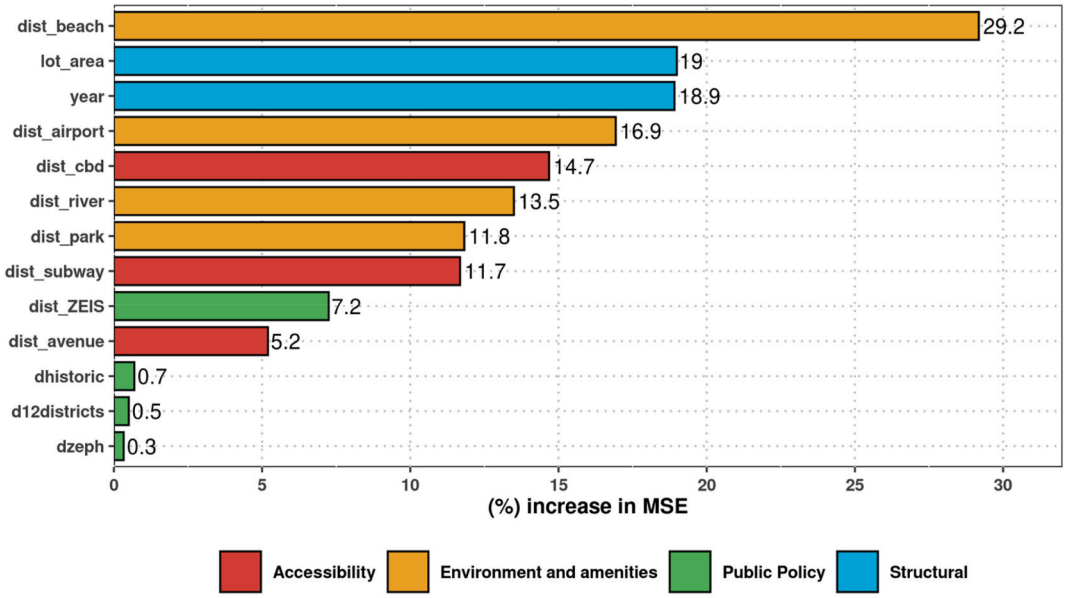


Fig. 3. Variable Importance - Increase in MSE (%) after permutation.

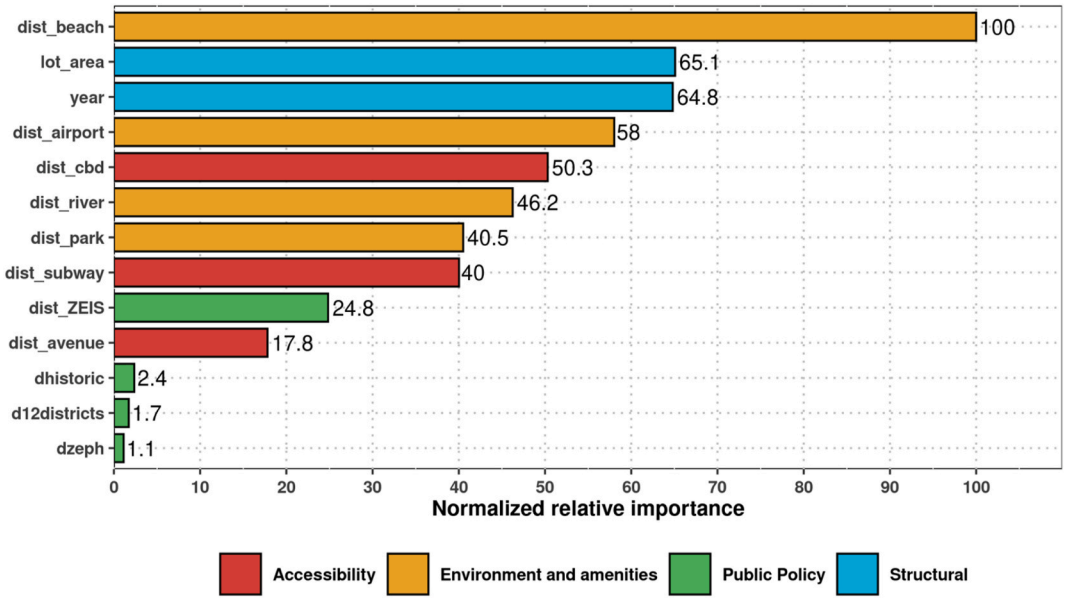
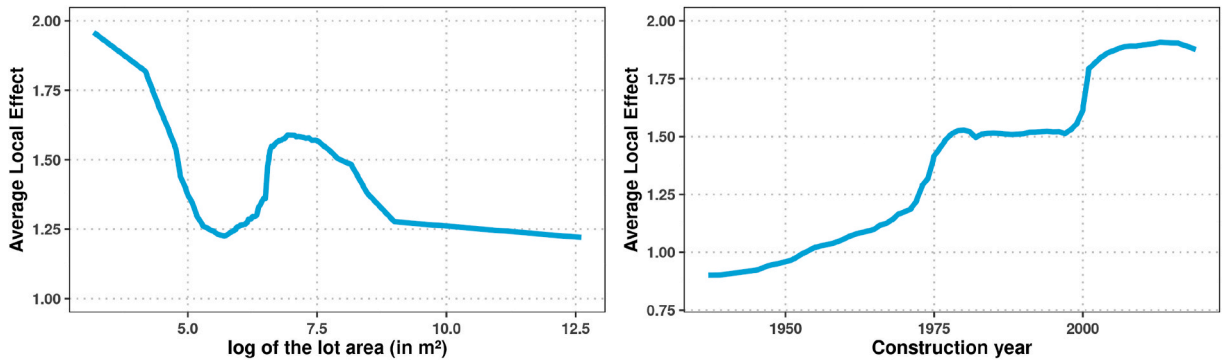


Fig. 4. Normalized relative importance of variables.



(a) (b)

Fig. 5. Accumulated Local Effect plots (ALE) plots for Structural covariates.

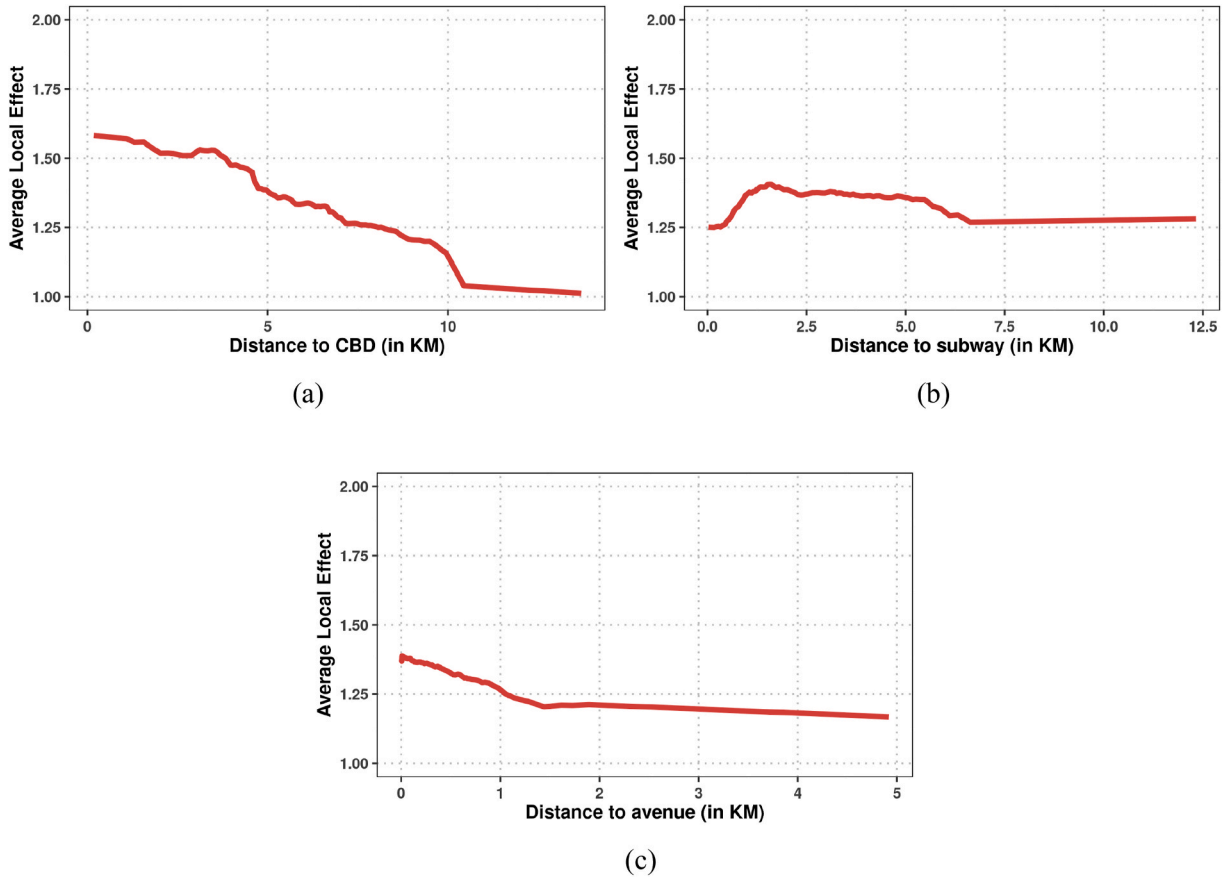


Fig. 6. Accumulated Local Effect plots (ALE) plots for Accessibility covariates.

also remains important for a wider range of distances.

Fig. 7(a)–7(d) indicate clear non-linear relationships between predicted FAR variations and the local values of the variables associated with amenities and the urban environment. The biggest variations in predicted FAR are related to the distance to the beach. Interestingly, these biggest FAR variations occur at short distances to the beach (up to 2 km), suggesting a very local amenity (something that may be associated with the absence of height restrictions for buildings near the beachfront). Thus, the most important FAR determinant in Recife appears to present stronger influences at short distances.

Finally, as indicated by Fig. 8, the FAR variations tend to increase at short distances to a ZEIS, suggesting that this kind of urban policy intervention negatively affects the value of lots closer to them. The relationship, however, is also non-linear: after 1 km, additional distance brings no additional FAR increases.

5. Discussion and conclusions

The more intensive use of urban land associated with the trend toward vertical housing is a current reality in Brazilian cities. Despite its importance for the quality of urban life, the subject is relatively little studied and the available evidence refers mainly to developed countries and was obtained using traditional linear multivariate regressions that do not adequately capture urban spatial heterogeneity (Barr, 2016; McMillen, 2006).

The results were obtained by considering a set of variables associated with accessibility to jobs and services, environmental and amenities factors, intrinsic lot characteristics, and urban land regulation on the FAR of multifamily residential buildings in Recife. These results indicate that these two ML techniques clearly outperform traditional OLS and LWR strategies and that RF outperforms XGBoost in predicting the FAR of multifamily residential buildings in the city. Using RF and the set of FAR determinants, it is possible to explain 62.5% of the variance FAR variance, a significant improvement considering, for example, the OLS regression result (about 34.7%). In addition, the evidence as a whole indicates that environmental and amenities factors, intrinsic lot characteristics (year and area), and accessibility indicators are all relevant to explain FAR behavior through the urban space. More specifically, the change in the urban environment and the amenities associated with the proximity to the beach is the most important factor underpinning the FAR variation across the urban sites of the city. Together with intrinsic lot characteristics, distance to the CBD is also still very relevant for understanding the FAR of lots in Recife. The third set of evidence involves the non-linear relationships between the FAR and its determinants. Very different from what is assumed in traditional empirical investigations about the FAR determinants but at the same time consistent with urban land heterogeneity, the results obtained using the ALE strategy indicated, for practically all the variables, the presence of non-linear effects on FAR.

Overall, these results bring different contributions to the available evidence about the determinants of FAR and have relevant policy implications. First, the evidence at the same time provides empirical support and qualifies the urban economics theoretical arguments for understanding the spatial urban variance of FAR. In this regard, the clear importance of factors associated with amenities (beach, river, park) and job accessibility (distances to CBD and subway stations) are in line with theoretical expectations (Brueckner et al., 1999; Fujita et al., 1989; Lee & Lin, 2018). The results are also in line with the evidence about the determinants of propriety prices and urban land uses in Recife (Lima & Silveira Neto, 2019; Seabra et al., 2016). Interestingly, the important association between distance to the beach and the value of FAR is in line with the relevance of natural amenities to explain the location of rich

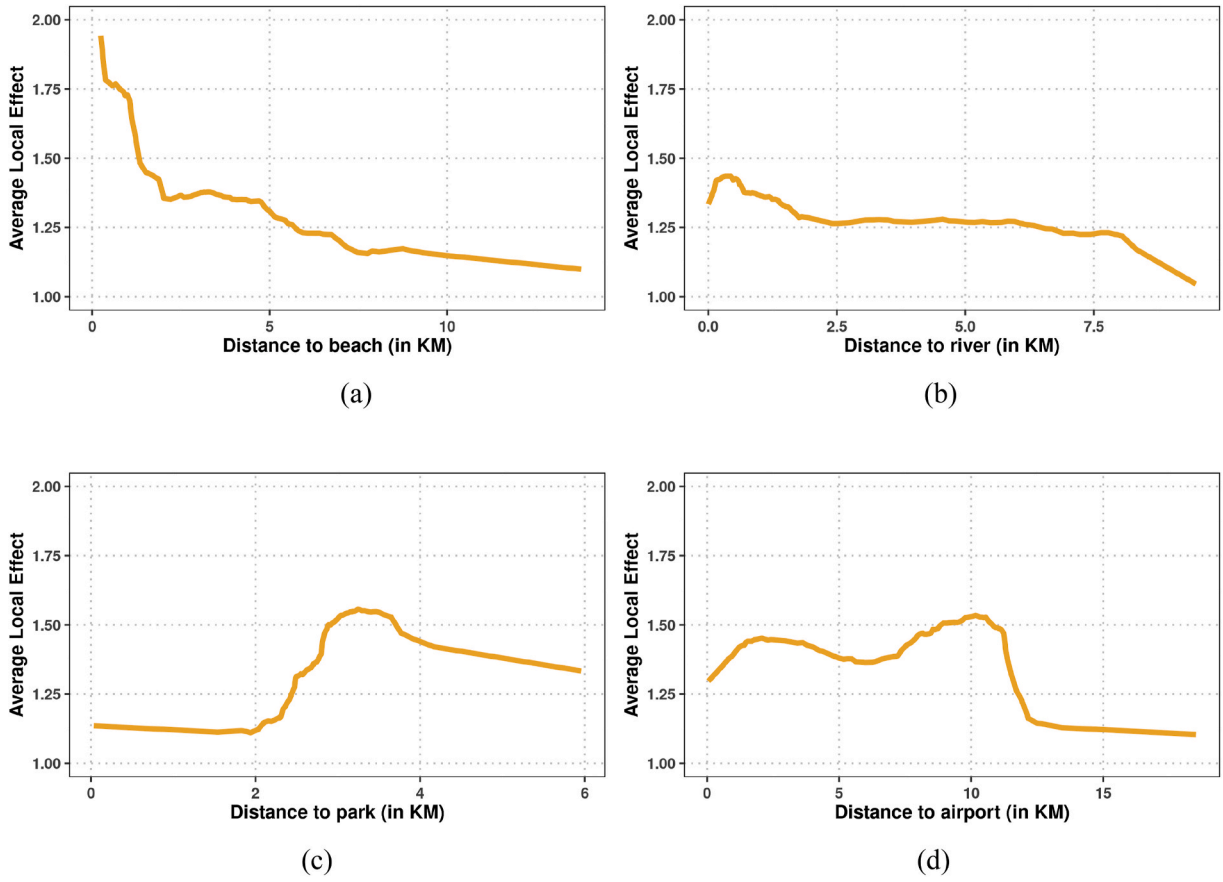


Fig. 7. Accumulated Local Effect plots (ALE) plots for Environment and amenities covariates.

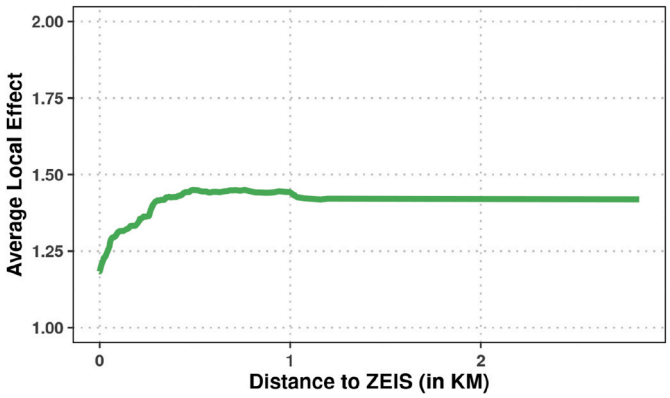


Fig. 8. The Accumulated Local Effect plot (ALE) for public policy covariate.

families in US cities found by Lee and Lin (2018). In a tropical city context, the relevance of proximity to the beach may even more strongly explain the social segregation of the city. Indeed, the importance of the CBD for the intensity of urban land use in Recife conforms very well to the historical monocentric nature of Recife, the most important initial Brazilian urban agglomeration linked to the international sugar trade.

But regarding the predictions of the theoretical models, we also found that important non-linearities are present in the effects of the variables on buildings' FAR. For example, as shown through the ALE analysis, the positive effects of proximity to the beach were highly important for buildings' FAR only up to approximately 2 km from the coast, behavior that seems entirely consistent with the strictly local character of this amenity. The effects of the area of the lots on their FAR, in turn, are negative for the smallest and biggest areas and positive for lots with intermediate sizes. This also seems consistent with the limitations for construction on the smallest lots and the loss of value of larger lots located on the periphery.

The set of evidence has clear urban policy implications. Traditionally, Brazilian municipal governments have tried to avoid or mitigate possible negative externalities associated with greater intensity in the use of urban land through restrictions on its use, a strategy that has raised housing prices and may contribute to the relegation of poorer people to the outskirts of the city (Dantas et al., 2018; Lima & Silveira Neto, 2019). Our evidence reveals the factors associated with the high intensity of urban soil in the city of Recife and important non-linearities present in these relationships. These information indicate the sites of the city that deserve special attention in terms of adequate provision of urban infrastructure and public services and the non-linearities found may generate more precision in the interventions since they indicate that the urban spaces are affected in specific ways by the determinants of the FAR. In this sense, in neighborhoods characterized by the presence of factors that generate the highest intensity of urban land use, the provision of a good public transport network and restrictions on car use, access to adequate sanitation and drinking water, and the guarantee of safety in public spaces are clear possible lines of actions for the city officials and planners. Rather than simply restricting urban land use, such policy directions can effectively contribute to mitigating potential amenities associated with more intensive urban land use. In addition, the non-linearities of the effects of the variables may indicate situations in which interventions can be spatially narrower or wider. For example, the evidence for the ALE indicates that policy interventions to mitigate the effect of more intense land use near the beach may be restricted to a maximum of 2 km from the shore, since from that distance the presence of the beach has reduced influence on the FAR of the lots.

In general, given the natural, social, and economic similarities, the

urbanization pattern present in the city of Recife is a good reflection of the prevailing conditions in the other coastal cities of the Brazilian Northeast. Thus, despite some differences in urban land regulation, our results can be useful for managers in other coastal cities of the region. Given the high level of residential segregation in Brazilian cities (Brueckner et al., 2019; Oliveira & Silveira Neto, 2015), following the perspective of Lee and Lin (2018), an important extension of the research should focus on understanding the roles of FAR determinants on the social segregation of the city of Recife. On the other hand, the results obtained for Recife are certainly less useful for understanding urban land use intensity in urban centers with fewer natural amenities or with a clearer polycentric structure (for example, Belo Horizonte or Porto Alegre). Thus, future research should consider other urban realities in Brazil. Particularly, for example, almost nothing is known about the determinants of urban land use intensity in the Midwest region of Brazil, where cities have grown the most in recent years.

Finally, an obvious limitation of the current study is its focus only on residential buildings. Despite being the majority of buildings, residential buildings have specific urban land use characteristics (Liu et al., 2018, 2020). Thus, a more comprehensive characterization of the urban land use intensity of a city demands a specific understanding of the determinants of the FAR of commercial lots and its variation according to economic activities (Rosenthal & Strange, 2019). This also should be part of a future investigation of the urban physical structure of Brazilian urban centers.

Author statement

The authors of the paper declare that the article was not published previously, that it is not under consideration for publication elsewhere, and that the publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out. They also agree that the work will not be published elsewhere in the same form, in English, or in any other language, including electronically without the written consent of the copyright holder.

Individual contributions to the paper are outlined below:

Célio Belmiro: Methodology, Software, and Writing.

Raul Silveira Neto.: Supervisions, Conceptualization, Writing, and Original draft preparation.

Andrews Barros: Methodology and Software.

Raydonal Ospina: Supervision, Writing-Reviewing, and Editing.

Declaration of competing interest

All authors of the paper declare that there are no financial and personal relationships with other people or organizations that could inappropriately influence (bias) their research. Therefore, they confirm that there are no potential competing interests related to employment, consultancies, stock ownership, honoraria, paid expert testimony, patent applications/registrations, and grants or other funding.

Acknowledgments

Support from the National Council for Scientific and Technological Development (CNPq – Bolsa de Produtividade) is gratefully acknowledged.

Appendix A. Correlation between variables

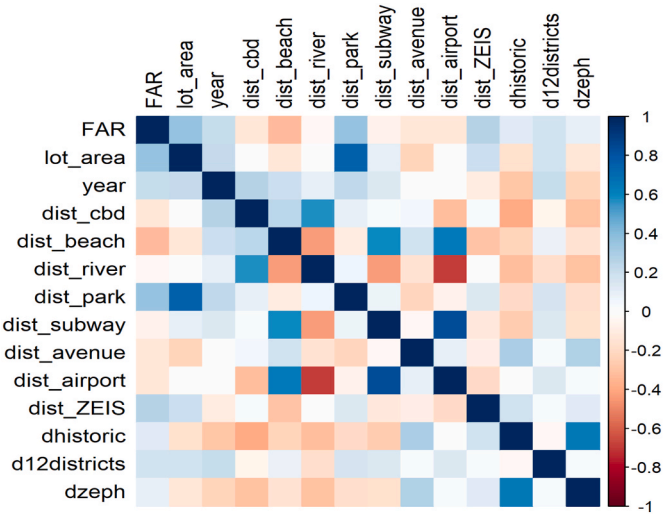


Fig. 9. Spearman correlation between variables

Table 3
Spearman correlation between variables

	FAR	lot_area	year	dist_cbd	dist_beach	dist_river	dist_park	dist_subway	Dist_avenue	dist_airport	dist_ZEIS	dhistoric	d12district	dzeph
FAR	1.00													
lot_area	0.36	1.00												
Year	0.22	0.23	1.00											
dist_cbd	−0.13	0.02	0.27	1.00										
dist_beach	−0.32	−0.13	0.19	0.25	1.00									

(continued on next page)

Table 3 (*continued*)

	FAR	lot_area	year	dist_cbd	dist_beach	dist_ river	dist_park	dist_ subway	Dist_ avenue	dist_ airport	dist_ ZEIS	dhistoric	d12district	dzeph
dist_river	−0.03	0.00	0.11	0.56	−0.44	1.00								
dist_park	0.36	0.75	0.24	0.09	−0.10	0.07	1.00							
dist_subway	−0.06	0.11	0.13	0.03	0.58	−0.43	0.07	1.00						
dist_avenue	−0.12	−0.23	0.01	0.04	0.19	−0.14	−0.22	−0.03	1.00					
dist_airport	−0.13	−0.01	0.00	−0.31	0.64	−0.70	−0.06	0.83	0.10	1.00				
dist_ZEIS	0.26	0.20	−0.11	0.02	−0.30	0.02	0.13	−0.12	−0.8	−0.19	1.00			
Dhistoric	0.12	−0.16	−0.29	−0.40	−0.23	−0.31	−0.20	−0.25	0.30	0.00	0.18	1.00		
d12district	0.18	0.19	0.22	−0.06	0.08	−0.17	0.16	0.14	0.03	0.14	0.03	−0.04	1.00	
Dzeph	0.10	−0.14	−0.22	−0.31	−0.15	−0.30	−0.17	−0.17	0.28	0.03	0.12	0.65	0.03	1.00

Appendix B. Scatterplot of actual and predicted values of FAR

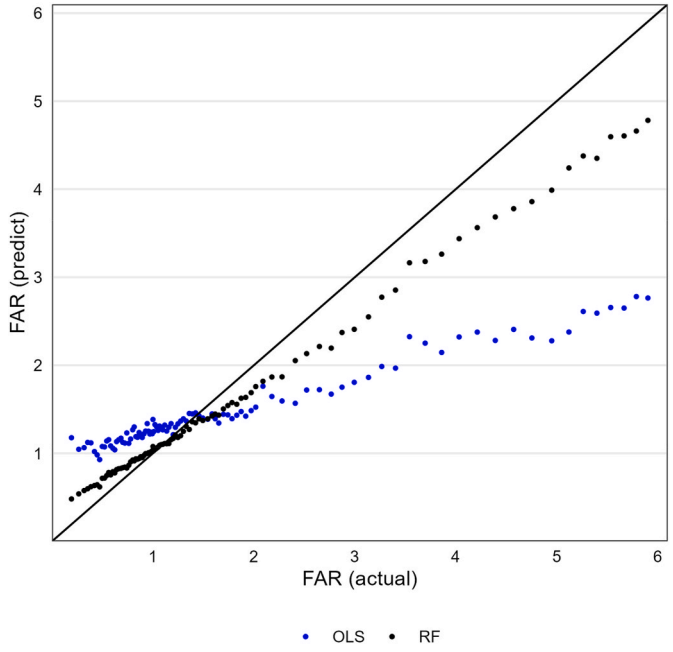


Fig. 10. Actual and predicted values of FAR for the OLS and Random Forest estimation. The values are for the testing samples only. To facilitate the visualization of relationships, graphs are constructed using average values of 100 bins grouping observations with the closest FAR values.

References

Abidoye, R. B., & Chan, A. P. (2017). Critical review of hedonic pricing model application in property price appraisal: A case of Nigeria. *International Journal of Sustainable Built Environment, 6*, 250–259.

Ahlfeldt, G. M., & Barr, J. (2022a). The economics of skyscrapers: A synthesis. *Journal of Urban Economics, 129*, Article 103419.

Ahlfeldt, G. M., & Barr, J. (2022b). Viewing urban spatial history from tall buildings. *Regional Science and Urban Economics, 94*, Article 103618.

Ahlfeldt, G., & McMillen, D. P. (2014). *New estimates of the elasticity of substitution of land for capital*. Louvain-la-Neuve: European Regional Science Association (ERSA).

Ahlfeldt, G. M., & McMillen, D. P. (2018). Tall buildings and land values: Height and construction cost elasticities in chicago, 1870–2010. *The Review of Economics and Statistics, 100*, 861–875.

Alfaro-Navarro, J. L., Cano, E. L., Alfaro-Cortés, E., García, N., Gámez, M., & Larraz, B. (2020). A fully automated adjustment of ensemble methods in machine learning for modeling complex real estate systems. *Complexity*, 1–12, 2020.

Apley, D. W., & Zhu, J. (2016). *Visualizing the effects of predictor variables in black box supervised learning models*. arXiv preprint arXiv:1612.08468.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys, 4*, 40–79.

Atkeson, C. G., Moore, A. W., & Schaal, S. (1997). *Locally weighted learning* (pp. 11–73). Lazy learning.

Aydinoglu, A. C., Bovkir, R., & Colkesen, I. (2021). Implementing a mass valuation application on interoperable land valuation data model designed as an extension of the national GDI. *Survey Review, 53*(379), 349–365.

Barr, J. (2012). Skyscraper height. *The Journal of Real Estate Finance and Economics, 45*, 723–753.

Barr, J. (2010). Skyscrapers and the skyline: Manhattan, 1895–2004. *Real Estate Economics, 38*, 567–597.

Barr, J. (2013). Skyscrapers and skylines: New York and chicago, 1885–2007. *Journal of Regional Science, 53*, 369–391.

Barr, J. (2016). The economics of skyscraper construction in manhattan: Past, present, and future. *International Journal of High-Rise Buildings, 5*, 137–144.

Barr, J., & Cohen, J. P. (2014). The floor area ratio gradient: New York city, 1890–2009. *Regional Science and Urban Economics, 48*, 110–119.

Belcher, R. N., Suen, E., Menz, S., & Schroepfer, T. (2019). Shared landscapes increase condominium unit selling price in a high-density city. *Landscape and Urban Planning, 192*, Article 103644.

Belmiro, C., Rodrigues, F., & Silveira-Neto, R. (2018). *Monocentrismo e estrutura urbana: uma análise empírica para a cidade do recife*. Rio de Janeiro, Brazil: Presented at 46° National Economics Meeting.

Bertaud, A., & Brueckner, J. K. (2005). Analyzing building-height restrictions: Predicted impacts and welfare costs. *Regional Science and Urban Economics, 35*, 109–125.

Bilgilioğlu, S. S., & Yılmaz, H. M. (2023). Comparison of different machine learning models for mass appraisal of real estate. *Survey Review, 55*(388), 32–43.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123–140.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Brueckner, J. K., et al. (1987). The structure of urban equilibria: A unified treatment of the muth-mills model. *Handbook of Regional and Urban Economics, 2*, 821–845.

Brueckner, J. K., Fu, S., Gu, Y., & Zhang, J. (2017). Measuring the stringency of land use regulation: The case of China's building height limits. *The Review of Economics and Statistics, 99*, 663–677.

Brueckner, J. K., Mation, L., & Nadalin, V. G. (2019). Slums in Brazil: Where are they located, who lives in them, and do they 'squeeze' the formal housing market? *Journal of Housing Economics, 44*, 48–60.

Brueckner, J. K., & Singh, R. (2020). Stringency of land-use regulation: Building heights in us cities. *Journal of Urban Economics, 116*, Article 103239.
Brueckner, J. K., & Sridhar, K. S. (2012). Measuring welfare gains from relaxation of land-use restrictions: The case of India's building-height limits. *Regional Science and Urban Economics, 42*, 1061–1067.
Brueckner, J. K., Thisse, J. F., & Zenou, Y. (1999). Why is central paris rich and downtown detroit poor?: An amenity-based theory. *European Economic Review, 43*, 91–107.
Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794). ACM.
Combes, P. P., Duranton, G., & Gobillon, L. (2011). The identification of agglomeration economies. *Journal of Economic Geography, 11*, 253–266.
Dantas, R. N., Duarte, G., Silveira Neto, R. M., & Sampaio, B. (2018). Height restrictions and housing prices: A difference-in-discontinuity approach. *Economics Letters, 164*, 58–61.
Danton, J., & Himbert, A. (2018). Residential vertical rent curves. *Journal of Urban Economics, 107*, 89–100.
Duranton, G., Henderson, V., & Strange, W. (2015). *Handbook of regional and urban economics*. Elsevier.
Duranton, G., & Puga, D. (2015). Urban land use. *Handbook of Regional and Urban Economics, 5*, 467–560. Elsevier.
Fan, J., Wang, X., Wu, L., Zhou, H., Zhang, F., Yu, X., Lu, X., & Xiang, Y. (2018). Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China. *Energy Conversion and Management, 164*, 102–111.
Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research, 15*, 3133–3181.
Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research, 20*, 1–81.
Florencio, L., Cribari-Neto, F., & Ospina, R. (2011). *Real estate appraisal of land lots using gamlss models*. arXiv preprint arXiv:1102.2015.
Friedman, J. H. (2001a). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.
Friedman, J. H. (2001b). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.
Fujita, M., et al. (1989). *Urban economic theory*. Cambridge Books.
Geshkov, M. V., & DeSalvo, J. S. (2012). The effect of land-use controls on the spatial size of us urbanized areas. *Journal of Regional Science, 52*, 648–675.
Glaeser, E. L., Gyourko, J., & Saks, R. (2005). Why is manhattan so expensive? Regulation and the rise in housing prices. *The Journal of Law and Economics, 48*, 331–369.
Gromping, U. (2020). Model-agnostic effects plots for interpreting machine learning models. *Reports in Mathematics, Physics and Chemistry*. Department II, Beuth University of Applied Sciences Berlin Report 1, 2020.
Gyourko, J., & Molloy, R. (2015). Regulation and housing supply. *Handbook of Regional and Urban Economics, 5*, 1289–1337. Elsevier.
Hong, J., Choi, H., & Kim, W.s. (2020). A house price valuation based on the random forest approach: The mass appraisal of residential property in South Korea. *International Journal of Strategic Property Management, 24*, 140–152.
Ho, W. K., Tang, B. S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research, 38*, 48–70.
Hu, L., He, S., Han, Z., Xiao, H., Su, S., Weng, M., & Cai, Z. (2019). Monitoring housing rental prices based on social media: An integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies. *Land Use Policy, 82*, 657–673.
Iban, M. C. (2022). An explainable model for the mass appraisal of residences: The application of tree-based machine learning algorithms and interpretation of value determinants. *Habitat International, 128*, Article 102660.
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning with applications in R*. New York: Springer.
Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F., & Ratti, C. (2021). Understanding house price appreciation using multi-source big geo-data and machine learning. *Land Use Policy, 111*, Article 104919.
Kunapuli, G. (2022). *Ensemble methods for machine learning*. Manning Publications.
Lee, S., & Lin, J. (2018). Natural amenities, neighbourhood dynamics, and persistence in the spatial distribution of income. *The Review of Economic Studies, 85*, 663–694.
Liaw, A., Wiener, M., et al. (2002). Classification and regression by random forest. *R News, 2*, 18–22.
Li, S., Jiang, Y., Ke, S., Nie, K., & Wu, C. (2021). Understanding the effects of influential factors on housing prices by combining extreme gradient boosting and a hedonic price model (xgboost-hpm). *Land, 10*, 533.
Lima, R. C. A., & Silveira Neto, R. M. (2019). Zoning ordinances and the housing market in developing countries: Evidence from Brazilian municipalities. *Journal of Housing Economics, 46*, Article 101653.
Lin, J., Zhuang, Y., Zhao, Y., Li, H., He, X., & Lu, S. (2022). Measuring the non-linear relationship between three-dimensional built environment and urban vitality based on a random forest model. *International Journal of Environmental Research and Public Health, 20*, 734.
Liu, C. H., Rosenthal, S. S., & Strange, W. C. (2018). The vertical city: Rent gradients, spatial structure, and agglomeration economies. *Journal of Urban Economics, 106*, 101–122.
Liu, C. H., Rosenthal, S. S., & Strange, W. C. (2020). Employment density and agglomeration economies in tall buildings. *Regional Science and Urban Economics, 84*, Article 103555.
McCluskey, W. J., Daud, D. Z., & Kamarudin, N. (2014). Boosted regression trees: An application for the mass appraisal of residential property in Malaysia. *Journal of Financial Management of Property and Construction, 19*, 152–167.
McMillen, D. P. (2006). Testing for monocentricity. *A companion to Urban Economics*, 128–140.
Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.
Murdoch, W., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences of the United States of America, 116*, 22071–22080.
Oliveira, T. G., & Silveira Neto, R. M. (2015). Segregação residencial na cidade do recife: Um estudo da sua configuração. *Revista Brasileira de Estudos Regionais e Urbanos, 9*, 71–92.
Pai, P. F., & Wang, W. C. (2020). Using machine learning models and actual transaction data for predicting real estate prices. *Applied Sciences, 10*, 5832.
Rosenthal, S., & Strange, W. (2019). *How close is close. The spatial reach of agglomeration economies*.
Seabra, D. M. S., Silveira Neto, R. M., & de Menezes, T. A. (2016). Amenidades urbanas e valor das residências: uma análise empírica para a cidade do recife. *Economia Aplicada, 20*, 143–169.
Taecharungroj, V. (2021). Google maps amenities and condominium prices: Investigating the effects and relationships using machine learning. *Habitat International, 118*, Article 102463.
Tchuente, D., & Nyawa, S. (2022). Real estate price estimation in French cities using geocoding and machine learning. *Annals of Operations Research*, 1–38.
Turner, M. A. (2005). Landscape preferences and patterns of residential development. *Journal of Urban Economics, 57*, 19–54.
Wassmer, R. W. (2006). The influence of local urban containment policies and statewide growth management on the size of United States urban areas. *Journal of Regional Science, 46*, 25–65.
Wei, P., Lu, Z., & Song, J. (2015). Variable importance analysis: A comprehensive review. *Reliability Engineering & System Safety, 142*, 399–432.
Wheeler, A. P., & Steenbeek, W. (2021). Mapping the risk terrain for crime using machine learning. *Journal of Quantitative Criminology, 37*, 445–480.
Yilmazer, S., & Kocaman, S. (2020). A mass appraisal assessment study using machine learning based on multiple regression and random forest. *Land Use Policy, 99*, Article 104889.
Zhou, Y., Huang, X., Chen, Y., Zhong, T., Xu, G., He, J., Xu, Y., & Meng, H. (2017). The effect of land use planning (2006–2020) on construction land growth in China. *Cities, 68*, 37–47.
Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B, 67*, 301–320.