Universitat de les Illes Balears

MASTER'S THESIS

MACHINE LEARNING CONTRIBUTIONS TO HEDONIC PRICING METHOD: ASSESSING HETEROGENEITY AND CAUSAL INFERENCE IN WILLINGNESS-TO-PAY

Llorenç Bartomeu Femenias Rosselló

Master's Degree in Anàlisi de Dades Massives en Economia i Empresa

(Specialisation/Pathway Eines en Gestió i Anàlisi Intel·ligent de Dades)

Centre for Postgraduate Studies

Academic Year 2020-2021

MACHINE LEARNING CONTRIBUTIONS TO HEDONIC PRICING METHOD: ASSESSING HETEROGENEITY AND CAUSAL INFERENCE IN WILLINGNESS-TO-PAY

Llorenç Bartomeu Femenias Rosselló

Master's Thesis

Centre for Postgraduate Studies

University of the Balearic Islands

Academic Year 2020-21

Key words:

Hedonic models, environmental valuation, Machine Learning, MWTP, flexibility, causal inference, heterogeneity, non-linearity.

Thesis Supervisor's Name: Dr. Jan Olof William Nilsson

Machine Learning contributions to Hedonic Pricing Method: assessing heterogeneity and causal inference in willingness-to-pay

Llorenç B. Femenias Rosselló
Tutor: Dr. Jan Olof William Nilsson

Treball de fi de Màster Universitari en Analisi de Dades Massives en Economia i Empresa (MADM)

Universitat de les Illes Balears
07122 Palma de Mallorca
Llorenc.femenias2@estudiant.uib.cat

Resumen

Los modelos de precios hedónicos constituyen uno de los métodos más extendidos a la hora de realizar ejercicios de valoración ambiental. Sin embargo, investigación previa ha identificado algunas limitaciones importantes en su aplicación: falta de flexibilidad en la definición de su forma funcional y falta de robustez en cuanto a su interpretación causal. Este trabajo propone la aplicación de algoritmos de Aprendizaje Automático (Machine Learning) para superar estas limitaciones y así proveer de estimadores más robustos referentes a la disposición marginal a pagar (MWTP por sus siglas en inglés) de los individuos por bienes ambientales de interés.

Abstract

Hedonic pricing models are one of the most widespread methods for conducting environmental valuation exercises. However, previous research has identified some important limitations in its application: lack of flexibility in the definition of its functional form and lack of robustness in terms of its causal interpretation. This work proposes the application of Machine Learning algorithms to overcome these limitations and, thus, to provide more robust estimators regarding the marginal willingness to pay (MWTP) of individuals for environmental goods of interest.

Key words: Hedonic models, environmental valuation, Machine Learning, MWTP, flexibility, causal inference, heterogeneity, non-linearity.

1. Introduction

1.1. Environmental Valuation

The environment provides well-being to society through the provision of essential goods and services for the proper functioning of the economy and, in general, to sustain life. While some of these goods and services, like natural resources such as wood, have an economic value recognized by society (are exchangeable in markets and, therefore, have a price), many others, that also generate social welfare, have an unknown value. Examples of the latter are natural areas that provide recreational services to society, such as beaches or natural parks. This peculiarity is due to their nature as public goods, that is, their consumption is non-exclusive and non-rival.

In this context, knowing the social preferences for public environmental amenities is essential to achieve proper environmental management.

Facing this problem, the field of environmental economics has developed a series of techniques known as Environmental Valuation Methods aimed to quantify, monetarily, the economic value of goods and services provided by ecosystems. These valuations are obtained through the analysis of individual and collective preferences for changes in ecosystems. That is to say, the economic valuation of environment studies how the welfare of society will be modified in the face of a change in the provision of environmental goods and services, and translates this shift in welfare into monetary units to, in a next stage known as project assessment, calculate the economic profitability of

environmental policies. Therefore, the final goal of the entire environmental valuation/assessment process is to achieve an optimal allocation of resources considering the preferences that society reveals for these non-market goods.

1.2. Hedonic Method: theory and limitations

One of the most extended models in the field of environmental valuation is the Hedonic Pricing Method, whose theoretical foundations are simple: in a specific housing market, buyers will choose which property to purchase depending on the attributes related to house structure (e.g., size, number of rooms) and to location (e.g., distance to green areas, views from the property, characteristics of the neighborhood) [1]. If the market works correctly each of these amenities will be capitalized within the price of the house. Once individuals observe the set of prices and characteristics associated with each property, buyers' purchases reveal what is their willingness-to-pay for marginal changes (MWTP) in property characteristics. For instance, through the application of this method, we can infer buyer's MWTP for each additional square meter, for each additional room or for each kilometer that the property approaches an environmental area (such as a beach or a green area).

In recent years, an important part of the specialized literature in this field has identified some limitations related to the assumption of linearity in the estimation of the Hedonic Pricing Valuation methods. Some examples are found in Kuminoff et al. [2], where it is noted that "*Our results suggest that large gains in accuracy can be realized by moving from the standard linear specifications for the price function to a more flexible framework*", or in Bishop et al. [3]: *"Theoretical and simulation evidence suggest that the hedonic price function should be assumed to be nonlinear (...) Semiparametric and nonparametric methods can provide additional flexibility in estimating hedonic price functions"*.

Increasing model flexibility would benefit environmental valuation allowing heterogeneity on MWTP. This heterogeneity could come from differences in MWTP for a specific attribute depending on the actual provision level of these attribute or depending on the provision level of another characteristic of the property. We exemplify these considerations through the case of the effect of beach distance on house prices: evidence suggests that distance does not have a linear effect on price since moving away in the first few kilometers (moving from the first coastal line to an interior area) should have a much greater effect on property's price than doing the same movement starting from a much more distant point. Even, reached a certain threshold, going further from the beach may not have any significant effect on housing prices [4]. Similarly, the effect of moving away from the beach may be different for properties with different characteristics, for example, the effect may be less in houses with a swimming pool than in those without.

Finally, the need to define methods that facilitate causal inference, as well as the mitigation of biases due to omitted variables, are also identified as opportunities to advance in the hedonic method literature. The implementation of quasi-experimental methods, closely related to the field of Program Evaluation Econometrics (e.g., Difference in Difference method or Propensity Score Matching), that study variations in the amenity of interest as a "treatment", are widely discussed in specialized literature [5].

1.3. Machine Learning to overcome these limitations

Simultaneously, in recent years, literature on the application of Machine Learning (ML) algorithms in economic research has grown exponentially [6, 7]. Following this line, and collecting the abovementioned limitations, we propose the application of ML techniques to provide greater flexibility to the pricing function as well as to improve the robustness of causal inference in the application of hedonic price modeling for environmental valuation. Specifically, we propose the training of several tree-based ML models (Random Forest and Gradient Boosting) and the analysis of their results using interpretability techniques (Partial Dependencies Plots) to determine whether MWTPs are linear for specific environmental amenities (section 4). This application can be included within the framework of on non-parametric hedonic modelling [8].

However, these methods can only solve the first of the exposed limitations, the non-linearity in the MWTP depending on the level of the amenity of interest itself. To increase the causal interpretation of the hedonic model and to assess heterogeneity in the MWTP of an environmental amenity depending on other characteristics of the property, we will apply causal inference ML techniques (Causal Forests, section 5). This method allows us to model a variation in the amenity of interest as if it were a treatment. Once applied, we can estimate the average effect of this treatment (comparable with the estimator obtained in the traditional hedonic linear regression model) and analyze whether this effect is heterogeneous depending on other housing

variables. This application can be included within the framework of quasi-experimental hedonic modelling [5].

We will apply all of the above techniques to a Melbourne's housing market dataset (analyzed in section 2). Among the variables included we find the sale price (dependent variable) as well as the main characteristics that determine house prices (surface area, number of rooms, age, ...). Finally, thanks to having the coordinates of each property included in the dataset, we have calculated the distance from each of these properties to each urban beach and to each green space in the city. These are our two amenities of interest and on which we will analyze the reported valuation (MWTP) for each model. Additionally, we will apply a hedonic linear regression model (section 3) whose estimators will serve as a baseline to evaluate the results obtained by ML models.

2. Data

The dataset used for this research is composed of 9.870 observations that came from the housing market of the Australian city of Melbourne in 2018 [1]. Each of these observations corresponds to a real state property for sale in the city.

2.1. Exploratory Analysis

Although the original data set contains 21 variables, some of them have been discarded due to their irrelevance in hedonic price modelling (e.g., property seller or property IDs).

We can divide the considered variables into three main groups depending on their nature [2]:

- Dependent variable: price (in Australian $) of the property sold.
- Physical characteristics of the house: nº of rooms, bedrooms, bathrooms and car spots, type of real state (house, townhouse or unit), land size (in m^2), building area (in m^2), year of construction
- Location: property's neighborhood (suburb), region, postal code, council area, latitude and longitude.

In Figure 1, we analyze the distribution of housing prices. As we can see, although it is true that the distribution is asymmetric and positive skewed (it presents a skewness coefficient greater than 2), we will not normalize the data to avoid losing interpretability in estimators [3].

To avoid problems related to multicollinearity (literature indicates that there may be strong correlations between house-related attributes), we have analyzed the correlation matrix between potential explanatory variables (Appendix B). The use of the variable "number of bedrooms" has been ruled out due to its strong correlation with other variables [4].

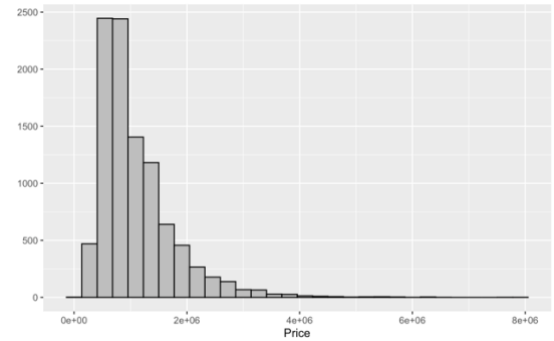

Figure 1: Price histogram.

Regarding the variables related to the property's location, it has been decided to work exclusively with "region" dummies. The following section discusses the role of geographic coordinates in the definition of new variables.

2.2. Geographical Analysis and Environmental amenities

Using the Python *GeoPandas (v 0.80)* library [5], each one of the properties included in the dataset has been projected from its coordinates. In figure 2, the geographical distribution of analyzed houses is mapped, differentiating by regions.

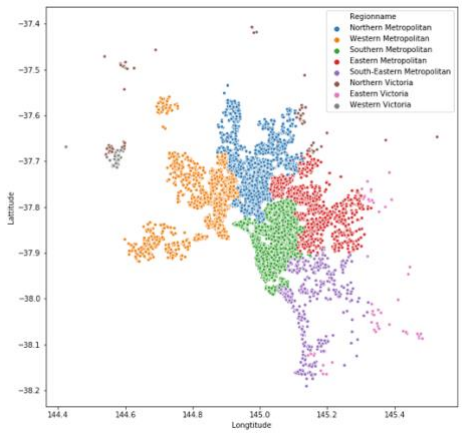

Figure 2: Geographic distribution of properties.

[1] The dataset is available on the Kaggle website: `https://www.kaggle.com/anthonypino/melbourne-housing-market`

[2] Section 3 delves into each of these groups

[3] Appendix A presents the main descriptive statistics of variables finally included in models.

[4] Correlation coefficient with "number of rooms" equal to 0.96

[5] https://geopandas.org

To ensure that we work with a unified and delimited market, we will include only those observations from the following regions: Northern, Western, Southern and Eastern Metropolitan [6]. Appendix C analyzes the existence of geographic patterns in the distribution of house prices in Melbourne.

Finally, it is important to comment on how the environmental variables of interest have been included. First, all urban beaches and urban green areas in the city of Melbourne have been geographically located. Once located, using the Python *Shapely (v. 1.7.1)* library [7], the distance (in km) from each house to each beach and urban green area has been calculated.

A large-size distance matrix between each property and each beach and green area has been obtained. From this matrix, we have extracted the distance from each house to the closest beach and green area. The distance to these locations is imputed as "Nearest beach" and "Nearest Green Area" variables. The distribution of these variables is presented in Appendix D.

Finally, it is important to note that all observations with missing values in any of the included variables have been discarded. Once this data cleaning process is finished, the final data set contains 8.360 observations.

3. Linear Hedonic Method

As a starting point, we are going to apply a traditional linear hedonic model estimated through OLS that will serve as a baseline to evaluate both the impact of Machine Learning models when identifying non-linearity in MWTPs, as well as to evaluate if the results obtained using ML methods are consistent with estimators obtained from traditional specifications. The latter will allow us to approximate the robustness of the flexible functional forms proposed in this research.

3.1. Methodology

As mentioned in the introductory section, the theoretical framework underlying hedonic price modeling is based on the identification of the values of the underlying characteristics of differentiated products through the observation of the market equilibrium price. In other words, if we examine the differences in the prices of two goods that differ in only one characteristic (e.g., two identical houses that only differ in the number of rooms), we can (indirectly) identify the tradeoffs (in terms of income willing to sacrifice for improvements in that characteristic) that individuals are willing to make regarding changes in that attribute.

In the specific case of environmental valuation, hedonic methods are mainly applied in housing markets. Starting from the observation of different prices and characteristics in sold houses, it is possible estimate the value that individuals obtain from these attributes. Following this premise, the choice of housing location is observable and directly related to environmental amenities of interest. If we put together all the previous points, we can conclude that the choice to purchase a property (and the price associated with it) implies a choice between different levels of environmental goods and services.

Partially deriving the hedonic price function on each property attribute, we can obtain the marginal price of each characteristic. Under certain conditions (that we will not consider due to the spatial delimitation of this research, for more information [1, 3]), these marginal prices are equal to the marginal willingness to pay (MWTP) for each house attribute. Finally, it is necessary to define the functional form that the Hedonic Price Function takes. While we want to use this section to establish the base line on which to compare the results obtained using innovative and flexible models, we will take as a reference the traditional linear function specification:

$$P = \alpha_0 + \sum_{i=1}^{h} \beta_i H_i + \sum_{j=1}^{n} \beta_j N_j + \sum_{k=1}^{l} \beta_k L_K + \varepsilon, \quad (1)$$

Where P is the (sale) price of a property; H represents physical characteristics of the house (number of rooms, land size…); N represents neighborhood characteristics; L represents location characteristics, such as proximity measures to areas of interest. It is in component L where the variables of distance to environmental amenities appear. Under this functional form, the implicit prices of any characteristic included in the function, z_i, is equal to the econometric parameter estimated for that attribute:

$$\frac{\partial P(z)}{\partial z_i} = \beta_i \ , \quad (2)$$

If we consider that the criteria by which we can match implicit prices with MWTPs are met, the estimators obtained from the econometric model represent MWTPs for each attribute. Estimators will be obtained through Ordinary Least Squares (OLS) method.

[6] The importance of the correct delimitation of the market being worked on is widely debated by Taylor, L.O. in chapter 10 of [1]

[7] https://pypi.org/project/Shapely/

3.2. Results

As noted in section 2, we will apply equation (1) with the following variables:

- P: price (in $) of the property sold
- H: number of rooms, number of bathrooms, building area, land size, type of property (house, unit or townhouse), year the house was built and number of car spots
- N: regional (4 categories) dummies
- L: distance (in km) to the nearest beach and to the nearest urban green area

These last two represent environmental amenities of interest. An important point to note is that we will use regional dummy variables that will serve as proxies for neighborhood characteristics. The reason that leads us to not consider suburb level, is that, although the suburb level should provide more precise information on the characteristics of the property's location, suburban level may have a greater correlation with environmental amenities [2, 9], leading to incorrect inference in estimators due to multicollinearity and spatial autocorrelation. Obtained results are presented in Table 1 [8].

Presented results fit those reported in previous literature: positive MWTPs for increases in the number of rooms, bathrooms, car spots, land size and building area are observed [10, 11, 12].

Focusing on environmental amenities, regressor estimators are negative for beach distance. In other words, for every kilometer that we move away from the beach, buyers' willingness to pay falls (or, if we put it the other way around, individuals show a positive willingness to pay to approach beaches). This means that beaches are environmental areas with a positive and significant impact in social welfare. MWTP for approaching one kilometer to the beach is $ 27,818 meaning that the price that individuals are willing to pay for a property decreases by more than $27,000 for each kilometer that the house is away from beach. In other words, for example, the price that the average buyer in this market will be willing to pay for a property located 1 kilometer away from the beach will be $ 27,818 lower than the price he/she would pay for an identical house on the beach shoreline. These results conform to those obtained in previous research [14, 15].

In the case of the Urban Green Areas, obtained results are very similar to those of beaches, however the effect is more moderate ($ 12,522 per kilometer). These findings are consistent with previous literature since MWTP is negative and significant [10, 13]. All coefficients are significantly different from 0 with a 99% confidence level.

Finally, it is important to comment that a R^2 close to 64% has been obtained, meaning that the model is capable of explaining about two thirds of the variance of the dependent variable [9].

	Dependent variable:
	Price
Nº Rooms	116,810.900***
	(7,668.298)
Type: Townhouse	−155,233.300***
	(18,561.350)
Type: unit	−339,251.600***
	(16,022.500)
Nº Bathrooms	206,645.200***
	(8,808.560)
Nº Car Spots	41,899.510***
	(5,289.803)
Land Size	29.512***
	(5.633)
Building Area	1,733.519***
	(67.787)
Construction Year	−3,794.469***
	(150.488)
Distance to beach	−27,818.590***
	(1,027.435)
Distance to Green Area	−12,522.260***
	(1,635.863)
Constant	8,076,952.000***
	(287,372.300)
Observations	8,360
R^2	0.637
Adjusted R^2	0.637
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 1: Linear Regression Estimators

4. Flexibilization of the price function through ML

As mentioned above, there exists empirical evidence that hedonic price functions in housing

[8] The estimators of the regional dummies (not included in the table) agree with the preliminary findings presented in Appendix C.

[9] In Appendix E, results for regression models with squared and cubical environmental amenities variables are presented and discussed. These models provide a greater degree of flexibility, but at the same time present a series of drawbacks

markets do not fit correctly to linear specifications [2, 3]. The solution to this problem lies in increasing model flexibility. Traditionally, this increase in flexibility has come through the application of non-parametric econometric approaches [8, 16, 17]. However, here we propose the use of non-parametric Machine Learning models that do not take any functional form or any relationship between variables beforehand.

4.1. Methodology

We will model housing price as a function of physical characteristics of the house (H), neighborhood dummies (N) and location characteristics (L). Specifically, the included location characteristics are, as mentioned above, distance to nearest beach and distance to nearest green urban area [10]:

$$P_i = f(H_i, N_i, L_i) + \varepsilon_i, \quad (3)$$

Within the wide variety of ML algorithms that can be used to solve regression problems, we will focus on tree-based methods. The main difference between these models and linear regression is that the former do not make assumptions about the functional form that is estimated. Furthermore, unlike linear methods that model the entire data set as a single function, tree-based models create a large number of learning subspaces that allow finding non-linear and monotic relationships and functions.

Tree-based ML methods' estimation strategy is based on the partitioning of the data in such a way as to minimize the sum of the squared errors (SSE):

$$SSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \text{ , (4)}$$

Where, in this study, N is the number of observations (properties), y_i is the true price of house *i* and $\hat{y}_i$ is the predicted house price. To increase accuracy, instead of training a single tree model, two popular ensemble methods have been used, *Random Forest* and *Gradient Boosting* (Appendix F). To avoid overfitting, the data set has been divided into a training set (75% of the observations) and a test set (the remaining 25%). In addition, a cross-validation process with 5 folds has been applied to find the combination of hyperparameters that maximizes predictive accuracy (Appendix G). Optimal hyperparameters and the metrics obtained by each model are presented in section 4.2.

[10] Some examples of literature that models housing prices from a Machine Learning can be found in Oladunni et al. [18], Park et al. [19] or Valier A. [20].

Although it is true that ML algorithms have a great predictive capacity, they do not stand out for their interpretability. This means that, on many occasions, these models can be seen as "black boxes" where the effect of the included predictors (features, in ML terminology) on outcomes cannot be observed. However, thanks to advances in the field of *interpretable machine learning*, this is starting to change [21, 22]. One of these techniques is known as *Partial Dependence Plot (PDP)* which is based on the graphical representation of the *Partial Dependency Function (PDF)* of each variable [23]. Thanks to this technique, we can see how the value of the predicted variable changes from marginal changes in the independent variable (predictor) of interest.

This technique has previously been used in environmental valuation studies based on hedonic models in Nafilyan et al. [24].

Mathematically, we can express it as follows: let $x = \{x_1, \dots, x_n\}$ represent the independent variables (predictors) in a ML model whose prediction function is $\hat{f}(x)$. If we select one of these predictors, z_i, and its compliment $z_c = x \setminus z_i$, we can define the *PDF* of the response on z_i as:

$$f_i(z_i) = E_{z_c}\left[\hat{f}(z_i, z_c)\right] = \int \hat{f}(z_i, z_c)\, p_c(z_c) d_{z_c} \text{ , (5)}$$

Where $p_c(z_c)$ is the marginal density function of (z_c),

$$f_i(z_i) = \int \hat{f}(z_i, z_c) \int p(x) d_{z_i}\, d_{z_c} \text{ , (6)}$$

Equation (5) can be estimated from training set by

$$\overline{f}_i(z_i) = \frac{1}{n} \sum_{j=1}^{n} \hat{f}\left(z_i, z_{j,c}\right) \text{ , (7)}$$

Where $z_{j,c}$ are the actual values of z_c in training set.

From an algorithmic point of view, we can summarize the construction of the partial dependency plot as

1. For $j \in \{1,2, \dots, k\}$
   (a) Copy the training data set and replace the original values of z_i with the constant $z_{i,j}$
   (b) Compute the vector of predicted values from the modified copy of the training set
   (c) Compute the average prediction to obtain $\overline{f}_i\left(z_{i,j}\right)$ from equation 7

2. Plot the pairs $\{z_{i,j}, \overline{f_l}(z_{i,j})\}$ for $j \in \{1,2,\dots,k\}$

Due to the special focus that this work has on the effect of certain environmental variables on the price of housing, once the partial dependence plots for these variables have been drawn, we will calculate the MWTP for each $z_{i,j}$ (where j represent each enter kilometer) taking the slope of the PDP at each point $z_{i,j}$ with respect to $z_{i,j-1}$ as represented in equation (8)

$$MWTP_{z_{i,j}} = \frac{\overline{f_l}(z_{i,j-1}) - \overline{f_l}(z_{i,j})}{z_{i,j} - z_{i,j-1}}, \quad (8)$$

In this way we will obtain a different MWTP for each level of the variables of interest. Specifically, we will calculate a flexible MWTP for each kilometer of distance with the environmental amenities of interest. Additionally, a linear regression model has been trained (following the same criteria for dataset division and cross-validation) that will help us to compare the prediction capacity of the ML models.

4.2. Results

ML models have been trained with the same set of variables that in section 3. After dividing the data set into a training and a test set, and applying a grid search with 5-fold cross validation (the hyperparameters selected for each model are presented in Appendix G), the following accuracy metrics have been obtained:

	RMSE		R2	
	Training (CV)	Test	Training (CV)	Test
Reg	424.373	366.867	0.643	0.632
RF	300.648	264.361	0.823	0.804
GBM	286.133	260.071	0.837	0.815

Table 2: ML Models Metrics

As we can see, ML non-parametric models achieve a greater predictive capacity (a RMSE reduction of 33%) and an increase in the explained variance (18 percentage points of improvement in R2) with respect linear regression[11]. In addition, it should be mentioned that no type of overfitting is seen in ML models since RMSE obtained for the test set is even smaller than in the training set (with cross validation)[12] . Regarding the importance of each variable (presented in Appendix I), we can see how the two environmental amenities included are important enough to be considered as relevant in the prediction and, therefore, in the hedonic price function.

Once the analysis of ML predictive capacity was carried out, we can proceed to analyze the effect of environmental variables on the property's price[13]

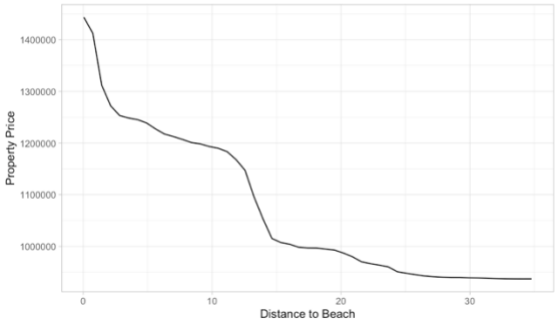

Figure 3: Distance to beach PDP

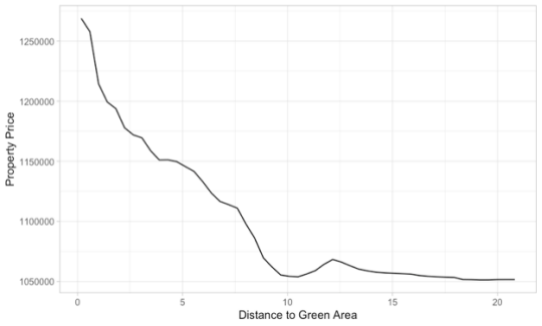

Figure 4: Distance to green area PDP

On the vertical axis we find the predicted property price. On the horizontal axis we plot the distance (in km) to the environmental amenity of interest. As we can see, in all cases, a strong non-linearity is observed. If we focus on the case of beaches, we can

[11] Appendix H includes accuracy plots

[12] This result may be due to the fact that once the cross-validation has been carried out, the model is retrained with the selected hyperparameters as well as with the entire training set, therefore the number of observations with which the model is trained increases.

[13] Due to the spatial limitations of this work, we only captured the PDPs for the Random Forest model. The PDPs obtained through Boosting are presented in appendix J. The reason for presenting the results of the RF model is due to the fact that it obtains predictive capacity metrics that are very similar to the GBM but PDP presents a smoother distribution.

see how both models find a very similar partial dependence function.

Partial Dependence Plot (Figure 3) seem to indicate that moving from the first coastline to the second one has a very large impact on the house price. After that, the effect moderates, with a range from 3 to 12 km where the slope of the function decreases. This slope becomes steeper again between kilometers 12 and 15, to moderate again between kilometers 15 and 25. Finally, we observe that from a threshold close to 25 kilometers distance, from which each additional kilometer has practically no effect on the price.

In the case of green areas (Figure 4), the effect does seem to be much more linear than in the beach case. The most remarkable output is that, approximately at a distance of 10 km, the price remains more or less constant. From these results we can infer that the effect of moving away from the beach is more persistent (it becomes null after 25 kilometers) and non-linear than in the case of green areas (whose effect disappears after 10 km).

Finally, following equation (8), we have calculated the MWTP for all levels (kilometers) of the environmental variables. The results are shown graphically in Figures 5 and 6.

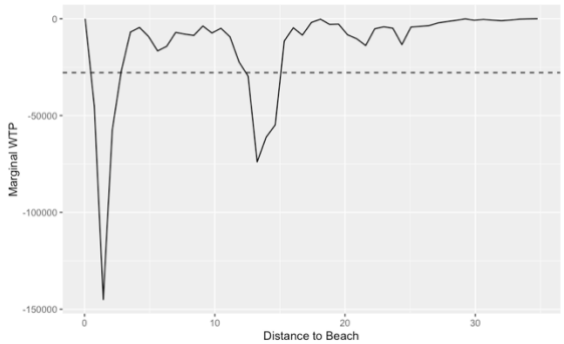

Figure 5: Distance to beach MWTP

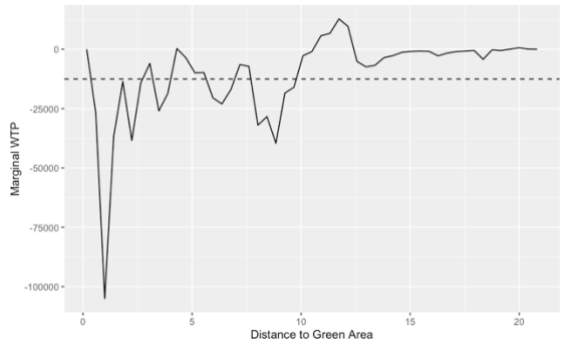

Figure 6: Distance to green area MWTP

The horizontal dashed line represents the MWTP obtained by the linear regression model in section 3, while the solid line graphs the result obtained by the Random Forest model. The results obtained are directly parallel to those presented in *PDPs* (Fig. 3 & 4).

In the case of beaches, we can see how going from the shoreline to 1 km away from beach implies a price drop of about A$ 150.000 (much higher than the A$ 27.818 estimated in the traditional regressive hedonic model). However, for the rest of the distances, the MWTP is close to 0. Nevertheless, a significant drop peak is detected at a distance of 12 km, with a MWTP of A$ 75.000. In the case of urban green areas, the observed dynamics are similar.

Therefore, the main conclusions of this section could be summarized as:

- ML models seem to achieve a better estimate of the real functional hedonic form since they achieve a lower prediction error than the linear regression model.
- A strong non-linearity is observed in the effect of the distance from the house to the environmental areas. Therefore, non-linearity is observed in MWTP.
- Large falls in the price of housing are observed when we move the first kilometer from the natural area of interest, but this effect disappears until it becomes close to zero at higher distances.

5. Improvement of causal inference and heterogeneity detection through ML

As pointed in Parameter & Pope (2013) [5], we can observe an increase in the number of papers that applies quasi-experimental designs in the field of hedonic valuation. This is the result of an increasing concern among researchers about the existence of biases due to omitted variables in the formulation of traditional hedonic models. Therefore, the causal interpretation of the results obtained may be questionable [3].

So far, the application of these quasi-experimental designs has been limited to estimating average treatment effects (ATE) equating them to the mean MWTP for specific attributes using, mainly, Difference-in-difference (DiD) [25, 26] and Regression Discontinuity (RD) [27] designs.

However, these models require either panel data (observations of the value of the same property before and after treatment) and the existence of an exogenous shock (e.g., the unforeseen construction of a highway that crosses a city) to estimate these MWTP. In order to avoid these specific data requirements, in this study we propose the application of a ML algorithm specially designed for

the extraction of causal inference known as Causal Forest.

In addition, this method allows us to face another of the existing limitations in hedonic models: the existence of heterogeneity in the MWTP of a specific attribute depending on its relationship with other attributes. In other words, this method allows us to study whether the MWTP of a specific attribute varies depending on the level of another attribute. This could be interpreted as the existence of Heterogenous Treatment Effects.

5.1. Methodology

Recently, there has been an increase in the development of Machine Learning models capable of obtaining causal interpretations from the implementation of quasi-experimental designs. Some of these first developments are found in works such as Athley & Imbens (2016) [28], where a method known as *Causal tree* is proposed. In this method, a partition of the dataset is made, with one of the parts used to build the regression tree while the other one is used to estimate the mean treatment effect within leaves.

As in tree-based ML methods, a consistent evolution that should improve model's accuracy (in this case, the predictability of treatment) is the use of ensemble methods. Thus, Athey et al. (2019) [29] presents the *Causal forest* method, which, as its name suggests, is based on Random Forest algorithm [14].

In our work, we will apply two *Causal Forest* models. In the first one, we will use the distance to the nearest beach as a "treatment" while, in the second one, the distance to the nearest green zone will substitute the former as “treatment”. Once the models have been trained, we can obtain the "Average Treatment Effect" (ATE) [30], comparable (at a theoretical level) to the MWTP estimator obtained by the traditional linear hedonic model. In this way, we will be able to evaluate whether the estimators obtained by the traditional hedonic model of section 3 are consistent with those obtained by a quasi-experimental design.

The dependent variable in both models continues to be the property price. Included explanatory variables (predictors) are the same than those included in sections 3 and 4. In addition, a grid search has been applied to find the optimal hyperparameter tuning.

One of the advantages of applying these models is that we can obtain "Heterogenous Treatment Effects" (HTE). The effect of the treatment can be extracted depending on the level of any other explanatory variable included in the model. In this way, we solve another of the limitations identified in the literature regarding hedonic modeling.

In the first place, we will evaluate which are the variables that present the greatest importance in the model training process, and then, by means of graphical plotting, evaluate the MWTP of the environmental variable of interest with respect to the level taken by other relevant explanatory variables (features or predictors).

5.2. Results

In this section we will analyze both the mean MWTP (as ATE) estimated by the quasi-experimental method (for the two environmental variables of interest), and the existence of heterogeneous effects in MWTPs.

5.2.1. Distance to nearest beach

Applying the causal forest model with the variable "distance to nearest beach" as treatment, we obtain an ATE (MWTP) of A$-31.127,339 (standard deviation equal to 1.395,679). As we can see, this estimate is slightly higher than the obtained by the traditional hedonic model (A$-27.818). Due to the reasonable similarity between the two, it seems that we are facing consistent results.

To study the existence of heterogeneous treatment effects, that is, the existence of differences in the MWTP for the distance to the beach depending on other characteristics of the house, we must analyze which are the variables with greater importance in the model. Results (for the 5 most important variable) are presented in Table 3.

Variable	Importance
Building Area	0.595
Nearest Green Area	0.114
Nº of Rooms	0.087
Nº of Bathrooms	0.063
Year of construction	0.033

Table 3: “Distance to beach” Causal Forest’s variable importance

Figure 7 plots the MWTP of the "distance to nearest beach" depending on the building area of the property.

[14] It is worth mentioning that other similar methods such as Causal Boosting and Causal MARS have been developed, however due to brevity they will not be studied nor applied in this work.

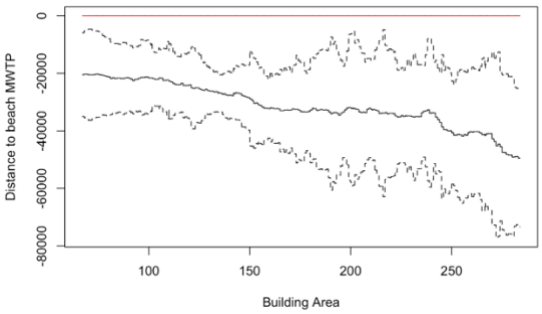

Figure 7: Distance to beach – Building Area HTE

Being the black continuous line the ATE (or MWTP in our case) and the dashed lines the 95% confidence interval, we can observe how the effect of moving one kilometer further from the beach increases (becomes more negative) as the size of the property increases. Although this result seems obvious, let us remember that the model already controls for the size of the built surface on the property. In other words, from these results we can extract that the distance to the beach has a greater effect on the price of large properties than in small ones.

This MWTP would range from approximately, \$-20.000 for each kilometer of distance to beach in properties with a built area of 50 m^2, to \$-40.000 in the case of built areas close to 300m^2.

Regarding the relationship between the MWTP of the distance to the beach with the other environmental amenity analyzed in this work, the results are presented in Figure 8.

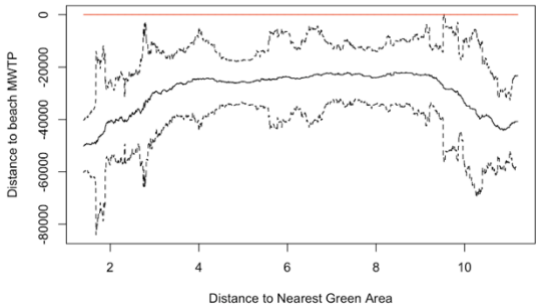

Figure 8: Distance to beach – Distance to Green Area HTE

In this case, heterogeneity in treatment is also observed. As we can see, the MWTP has an inverted U shape: moving away from the beach has more negative effects on the price that buyers are willing to pay when this property is either very close to or far from a urban green area.

In this case, the MWTP ranges from \$-50.000 at the most extreme points to \$-30.000 at intermediate distances (4 to 9 km) to green areas.

It is important to note that, in both cases (Figures 7 and 8), the results are significantly different from 0 for all levels. But this does not have to be the general rule in all cases. For example, Figure 9 shows the MWTP of the distance to the beach depending on the year the house was built.

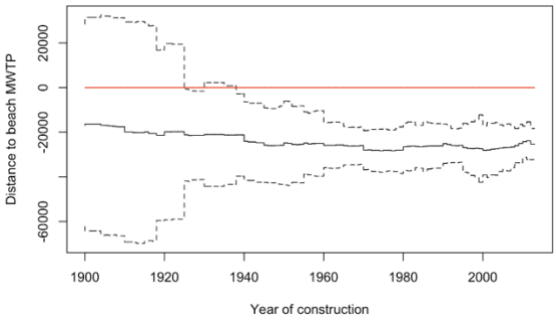

Figure 9: Distance to beach – Year of construction HTE

In this example, we can see that for older homes the distance to the nearest beach does not have any significant effect on the property's price.

5.2.2. Distance to nearest green area

If we use the variable "distance to nearest green area" as a treatment, we obtain an ATE equal to \$-16.808,09 (standard error equal to 2.628,35). As we can see, as in the case of beach distance, this estimate is slightly higher than that obtained by the traditional linear hedonic model (\$-15.522). Therefore, these results seem consistent with those obtained by traditional methods.

As in the previous section, the variables that present the greatest importance in the model are analyzed below.

Variable	Importance
Southern-Metropolitan	0.448
Building Area	0.220
Nearest Beach	0.141
Land Size	0.069
Nº of Bathrooms	0.042

Table 4: "Distance to green area" Causal Forest's variable importance

As we can see, some of the variables do coincide with the previous case. Due to the dummy nature of the regional identification variable, we will not

analyze whether there is heterogeneity in the MWTP depending on it.

Figure 10 plots the relation between the MWTP of the variable "distance to nearest green area" and the size of the building area of the property.

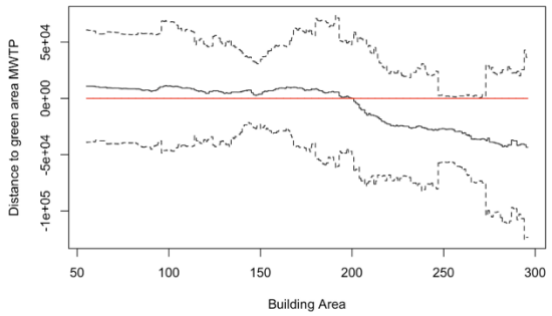

Figure 10: Distance to green area – Building Area HTE

However, unlike the case of distance to the beach, there does not appear to be significance in the MWTP of this environmental attribute at any level of the Building Area. A similar result is obtained if we analyze the relationship between the distance to the beach and the MWTP for green areas (Figure 11).

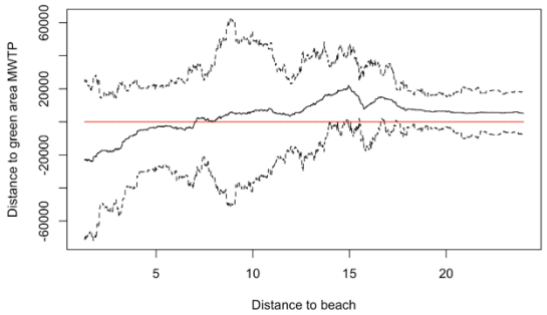

Figure 11: Distance to green area – Distance to beach HTE

6. Conclusions

As discussed throughout this essay, this project aims to provide new methodologies to solve two relevant limitations identified in the field of hedonic price modeling for environmental valuation.

These limitations are related to the need to provide greater flexibility to the price function (thus allowing the identification of heterogeneity in the willingness to pay for environmental attributes depending on the level of the attribute itself as well as other variables of the price function) and to the need to increase the causal interpretability of the models.

Traditional responses to these limitations have been the application of non-parametric econometric methods (in order to increase hedonic price function flexibility) and the design of quasi-experiments (to increase the causal inference reliability).

The main contribution of this work is the application of techniques based on Machine Learning to continue the path of overcoming the aforementioned limitations. In order to empirically test whether these techniques provide satisfactory results, a dataset of the housing market in Melbourne has been used. The environmental ecosystems evaluated were the beaches and urban green areas of this city.

In the first place, it has been verified through the application of a traditional hedonic model (through linear regression) how both environmental attributes are considered as goods for the individuals that participate in this market, in so far as willingness to pay for properties decreases as we move away from beaches and green areas (negative MWTP).

Secondly, Partial Dependence Functions of various tree-based ML models have been analyzed in order to obtain the relationship between both environmental attributes and housing prices. From this analysis, MWTPs has been derived for all levels of both environmental variables, revealing a strong non-linearity. Therefore, these results fit with those identified in previous literature and seem to reveal that these techniques may represent a potential alternative to traditional methods.

Finally, using these environmental variables as if they were treatments, a *Causal Forest* model has been applied to our dataset. Thanks to this model it has been possible to identify both the mean MWTP for each environmental attribute and the presence of heterogeneity in the MWTP depending on the levels of other housing attributes.

Using this technique, a similar, but slightly higher, MWTP has been obtained for both environmental attributes than in linear model. In addition, the existence of heterogeneity in MWTP for environmental variables depending on other property's characteristics has been confirmed.

Although it is true that promising results have been obtained, there are still many limitations in the application of these techniques for environmental valuation. Without the intention of being exhaustive, some of these limitations are: need to evaluate if the actual MWTP is obtained from the derivation of the Partial Dependence Function (or if we could only speak of capitalization) or increase the number of controls that allow ensuring the reliability of the quasi-experimental design.

References

[1] Peterson, L. G. (2003). *A primer on nonmarket valuation* (Vol. 3). P. A. Champ, K. J. Boyle, & T. C. Brown (Eds.). Dordrecht: Kluwer Academic Publishers.

[2] Kuminoff, N. V., Parmeter, C. F., & Pope, J. C. (2010). Which hedonic models can we trust to recover the marginal willingness to pay for environmental amenities?. *Journal of environmental economics and management*, *60*(3), 145-160.

[3] Bishop, K. C., Kuminoff, N. V., Banzhaf, H. S., Boyle, K. J., Pope, J. C., Smith, V. K., ... & von Gravenitz, K. (2019). Best Practices in Using Hedonic Property Value Models for Welfare Measurement. *Review of Environmental Economics and Policy*, *43*.

[4] Landry, C. E., & Hindsley, P. (2011). Valuing beach quality with hedonic property models. *Land Economics*, *87*(1), 92-108.

[5] Parmeter, C. F., & Pope, J. C. (2013). Quasi-experiments and hedonic property value methods. In *Handbook on experimental economics and the environment*. Edward Elgar Publishing.

[6] Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3-28

[7] Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87-106.

[8] Parmeter, C. F., Henderson, D. J., & Kumbhakar, S. C. (2007). Nonparametric estimation of a hedonic price function. *Journal of Applied Econometrics*, *22*(3), 695-699.

[9] Basu, S., & Thibodeau, T. G. (1998). Analysis of spatial autocorrelation in house prices. *The Journal of Real Estate Finance and Economics*, *17*(1), 61-85.

[10] Morancho, A. B. (2003). A hedonic valuation of urban green areas. *Landscape and urban planning*, *66*(1), 35-41

[11] Jim, C. Y., & Chen, W. Y. (2009). Value of scenic views: Hedonic assessment of private housing in Hong Kong. *Landscape and urban planning*, *91*(4), 226-234.

[12] Hansen, J. (2009). Australian house prices: a comparison of hedonic and repeat-sales measures. *Economic Record*, *85*(269), 132-145.

[13] Saphores, J. D., & Li, W. (2012). Estimating the value of urban green areas: A hedonic pricing analysis of the single family housing market in Los Angeles, CA. *Landscape and urban planning*, *104*(3-4), 373-387.

[14] Gopalakrishnan, S., Smith, M. D., Slott, J. M., & Murray, A. B. (2011). The value of disappearing beaches: a hedonic pricing model with endogenous beach width. *Journal of Environmental Economics and Management*, *61*(3), 297-310.

[15] Landry, C. E., Turner, D., & Allen, T. (2019). Hedonic property prices and coastal beach width. *Available at SSRN 2474276*.

[16] Mason, C., & Quigley, J. M. (1996). Non-parametric hedonic housing prices. *Housing studies*, *11*(3), 373-385.

[17] McMillen, D. P., & Redfearn, C. L. (2010). Estimation and hypothesis testing for nonparametric hedonic house price functions. *Journal of Regional Science*, *50*(3), 712-733.

[18] Oladunni, T., & Sharma, S. (2016, December). Hedonic housing theory—a machine learning investigation. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 522-527). IEEE.

[19] Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, *42*(6), 2928-2934.

[20] Valier, A. (2020). Who performs better? AVMs vs hedonic models. *Journal of Property Investment & Finance*.

[21] Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, *63*(1), 68-77.

[22] Zhao, Q., & Hastie, T. (2021). Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, *39*(1), 272-281.

[23] Greenwell, B. M. (2017). pdp: An R Package for Constructing Partial Dependence Plots. *R J.*, *9*(1), 421.

[24] Nafilyan, V., & Lorenzi, L. Valuing green spaces in urban areas: a hedonic price approach using machine learning techniques.

[25] Winke, T. (2017). The impact of aircraft noise on apartment prices: a differences-in-differences hedonic approach for Frankfurt, Germany. *Journal of Economic Geography*, *17*(6), 1283-1300.

[26] Boes, S., Nüesch, S., & Wüthrich, K. (2015). Hedonic valuation of the perceived risks of nuclear power plants. *Economics letters*, *133*, 109-111

[27] Sue, E. D., & Wong, W. K. (2010). The political economy of housing prices: Hedonic

pricing with regression discontinuity. *Journal of Housing economics*, *19*(2), 133-144.
[28] Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353-7360.
[29] Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *Annals of Statistics*, *47*(2), 1148-1178.
[30] Heckman, J. J., & Vytlacil, E. J. (2007). Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation. *Handbook of econometrics*, *6*, 4779-4874.

Appendix A. Descriptive Statistics

Statistic	Mean	St. Dev.	Min	Max
Price	1,098,910.000	682,586.000	85,000	8,000,000
Rooms	3.035	0.977	1	10
Bathroom	1.623	0.715	0	9
Car	1.655	0.967	0	10
Landsize	504.301	897.578	0	37,000
BuildingArea	148.689	89.063	0.000	3,112.000
YearBuilt	1,965.090	37.215	1,196	2,019
nearest_beach	10.239	6.660	0.051	34.786
nearest_green_area	5.447	3.234	0.173	20.823

Table 5: Descriptive Statistics of explanatory variables

The average price of properties for sale is close to A$1.1 million. The average home dates back to 1965 and is characterized by 3 bedrooms, between 1 and 2 bathrooms (average 1.62) and between 1 and 2 car spots (average 1.65). The average property size is 504 m^2 with an average built-up area of 148 m^2.

The average distance to the nearest beach is 10.2 kilometers. In this variable there is a great variation among the different properties as the distance varies from just over 50 meters to nearly 34 kilometers. Similar results are obtained in the case of the distance to the nearest green area, with an average of 5.4 kilometers. In this case, there is less variation between properties.

Appendix B. Correlation Matrix

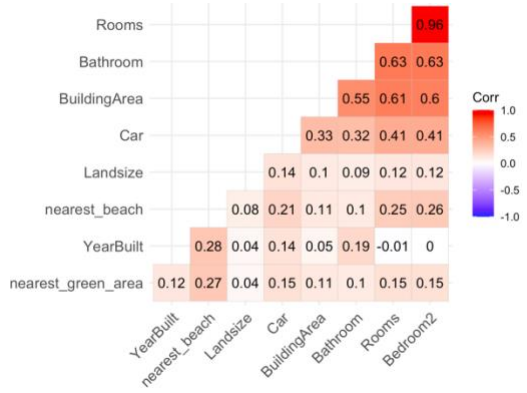

Figure 12: Correlation Matrix

As we can see at a glance, there is a strong correlation between the number of rooms and the number of bedrooms in a house. These results are not surprising and are in line with those published in other similar studies.

Due to the fact that the remaining variables have a correlation of less than 0.7 (a value that is usually taken as a reference when evaluating the potential presence of multicollinearity), it was decided to eliminate only the variable "Bedroom" from our study.

Appendix C. Geographical Patterns of Price Distribution

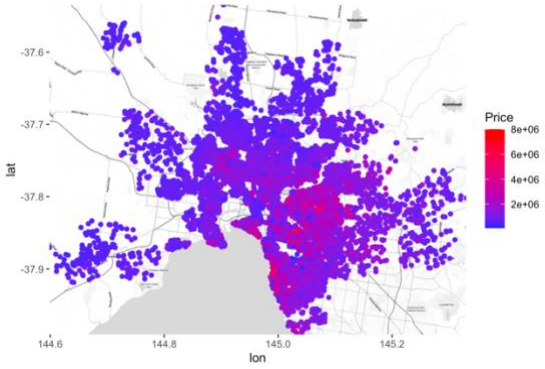

Figure 13: Geographical distributions of housing prices in the city of Melbourne

In figure 13 we can see how there are clear patterns that justify the inclusion of location dummy variables (suburb or region). Concentrations of high prices are observed in the "Southern Metropolitan" region.

Furthermore, it can be observed how, a priori, there is an increase in prices in areas near the coast as well as around green areas (which can be identified from large gaps without observations on the map). These preliminary observations must be confirmed in the models that will be presented in next sections.

Appendix D. Environmental Amenities Distribution

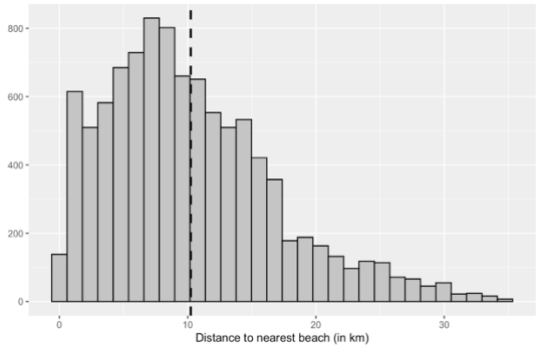

Figure 14: Distance to beach histogram.

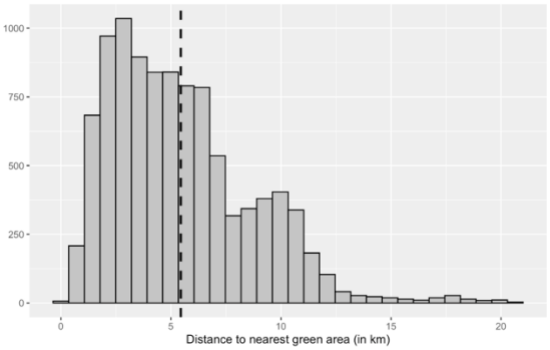

Figure 15: Distance to green area histogram.

In both cases we find variables with a quasi-normal distribution since they have a low skewness (skewness coefficient close to or less than 1).

Appendix E. Ampliation of Linear Regression Hedonic Model

One of the more traditional ways of adding flexibility to hedonic pricing functions has been by including the squared (and higher degree) factors of the interest amenities. In Table 6, we present the results of the re-estimated OLS regression including the squared and cubical terms for the environmental variables studied.

Following previous literature results we expect a negative coefficient for standard environmental variables and a positive estimate for the squared terms. These results imply that the negative effect of moving away from environmental amenities reaches a maximum point in a certain threshold. Above this threshold this negative effect starts to disappear.

The coefficients of the "Green Area" variables meet these expectations (col. 1 in Table 6). However, the same does not happen with the distance to nearest beach. According to the coefficients presented, the effect of moving away from beach presents an increasing (more negative) MWTP with distance.

The interpretation of the cubic factors model (col. 2 in Table 6) is very tricky if we just observe regression coefficients. To facilitate this understanding, we are going to graph the MWTP for each km in figures 16 and 17.

	Dependent variable:	
	Price	
	(1)	(2)
N° Rooms	117,465.800***	123,715.500***
	(7,656.873)	(7,610.425)
Type: Townhouse	−159,749.700***	−176,527.900***
	(18,552.510)	(18,454.760)
Type: unit	−344,009.500***	−354,089.600***
	(15,985.680)	(15,874.320)
N° Bathrooms	206,422.300***	202,714.200***
	(8,802.398)	(8,742.735)
N° Car Spots	42,620.710***	44,432.170***
	(5,287.923)	(5,245.976)
Land Size	28.833***	29.102***
	(5.615)	(5.569)
Building Area	1,732.173***	1,698.434***
	(67.587)	(67.078)
Construction Year	−3,700.892***	−3,412.704***
	(150.490)	(151.148)
Distance to beach	−18,370.650***	17,399.870***
	(2,596.994)	(5,353.901)
Squared Distance to beach	−386.356***	−3,824.223***
	(88.770)	(454.739)
Cubical Distance to beach		79.495***
		(10.843)
Distance to Green Area	−45,649.090***	33,692.580***
	(4,717.352)	(11,146.480)
Squeared Distance to Green Area	2,450.586***	−8,380.402***
	(324.193)	(1,529.904)
Cubical Distance to Green Area		380.805***
		(60.675)
Constant	1 7,916,855.000***	7,201,192.000***
	(287,416.200)	(291,285.900)
Observations	8,360	8,360
R^2	0.640	0.646
Adjusted R^2	0.639	0.645
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 6: Linear Regression Estimators (with squared terms)

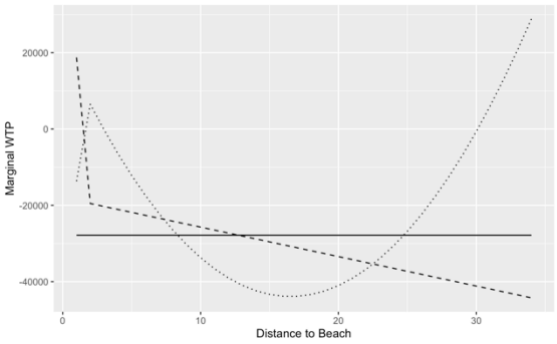

Figure 16: Distance to beach MWTP

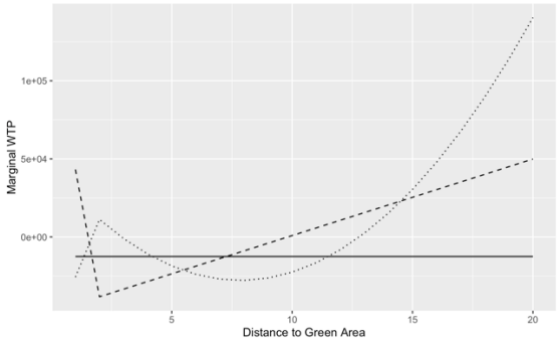

Figure 17: Distance to Green Area MWTP

Solid line represents MWTP for linear model (Table 1), dashed line represents squared model MWTP and dotted line represents MWTP for cubical model in Figures 16 and 17.

As we can see, the linear model presents a constant MWTP (represented by the solid horizontal line) for all distances. However, the squared model (dashed line) presents two different forms depending on the variable analyzed: in the case of the distance to green areas it does comply with the expected v-shaped effect, however for the distance to beaches variable this form is not fulfilled, and we observe an elbow-shaped line with an increasingly negative effect. In the case of the cubic model (dotted line) we observed the same pattern for the two variables analyzed.

Appendix F. Random Forests and Gradient Boosting

Tree-based methods are among the most used in the field of Machine Learning. In the case of regression problems (where the variable to be predicted takes continuous values), the algorithm partitions a data set into multiple subsets following a series of rules that act recursively on a data set. Each selection rule divides the data into smaller sets so that the sum of squared residuals is minimized (equation 4). Then the value that the tree predicts for each observation in a region determined by these rules (terminal nodes) will be the average of the response values of the training observations that fall in this region.

To increase the prediction accuracy of these models, multiple ensembling techniques have been developed. One of these is the Random Forest algorithm. This algorithm, which can be considered an evolution of Bagging techniques, aims to reduce the variance generated in the predictions of ML models.

Given a set of n independent observations $(Z_1, \dots, Z_n)$ with variance σ^2, the variance of the mean of these observations, $\bar{Z}$, will be $\frac{\sigma^2}{n}$ or, more commonly noted, $\frac{\sigma}{\sqrt{n}}$.

Therefore, if we obtain a sample of predictions generated from different subsets of data (generating different trees) for each observation, we can reduce the variance in the model predictions. To do this, a boostraping technique is applied on the training set data (generating B subsets of the training set) and we obtain the prediction for each observation from the mean of the predictions of each of these B sets. Being the final prediction reported by the model for observation *i* with the set of *x* explanatory variables:

$$\hat{y}_i = \hat{f}_i(x) = \frac{1}{B}\sum_{b=1}^{B} \tilde{f}_i^b(x)$$

Where $\tilde{f}_i^b(x)$ is the predicted value for observation *i* with explanatory variables *x* in the boostrap set *b.*

On this basis, the Random Forest algorithm eliminates the potential correlation between trees (which can lead to an increase in variance) because each tree is constructed considering only a random selection of *m* predictors, chosen randomly from the entire set of *p* predictors.

The other algorithm considered in this work is Gradient Boosting. This algorithm works in a similar way to those related to bagging techniques (such as random forests) presented above, except that in this case the trees grow sequentially, that is, each tree grows taking into account the information derived from the tree that precedes it. In this way, the algorithm "stores" the information of the residues obtained in each trained tree sequentially, thus passing the information to the next stages of the training process.

Appendix G. Cross-Validation, Metrics and Hyperparameter Tuning

Following the traditional recommendations to avoid overfitting in the models, a cross-validation method with 5 folds has been applied. That is, the training set has been randomly divided into 5 subsets and the models have been recursively trained with 4 of these subsets, leaving one as the validation set.

As we are faced with a regression problem, the metric used to evaluate the precision of the model has been the Root Mean Squared Error (RMSE), which is calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\sqrt{(y_i - \hat{y}_i)^2}}{N}}, (X)$$

Where N is the number of observations, y_i is the true price of house i and $\hat{y}_i$ is the prediction house price.

Finally, a Grid-Search process (together with cross validation) has been applied to obtain those optimal hyperparameters. For each model, a search was carried out among the following parameters:

- Random Forest: nº of randomly selected predictors for each tree (2, 5, 6, 8 or 10)
- Gradient Boosting: interaction depth (20, 25 or 30), nº of trees (500, 2000, 3500, 5000 pr 6500) and shrinkage (0.01 or 0.001).

The following figures graphically show the RMSE obtained during the cross-validation training process for each combination of hyperparameters.

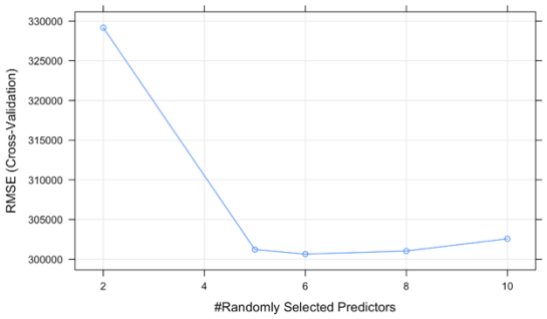

Figure 18: Random Forest Grid-Search

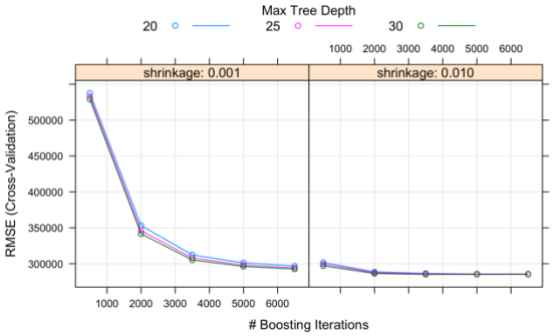

Figure 19: Boosting Grid-Search

Finally, the combination of hyperparameters that the RMSE has minimized in the cross-validation process for each model has been:

- Random Forest: 6 randomly selected predictors for each tree
- Gradient Boosting: interaction depth equal to 30, 2000 trees and a 0'01 shrinkage.

Appendix H. Prediction accuracy

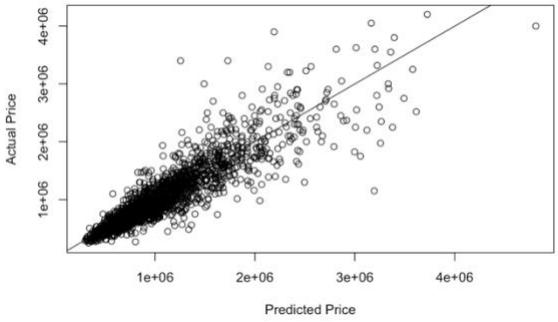

Figure 20: Random Forest Accuracy

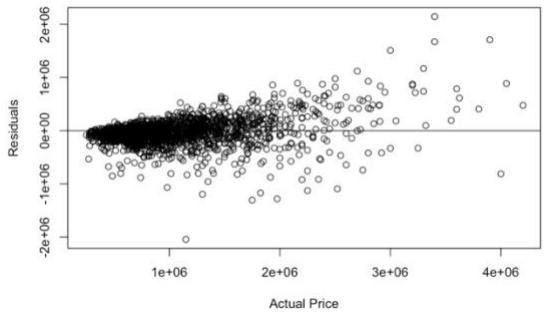

Figure 21: Random Forest Residuals

Appendix I. Variable Importance

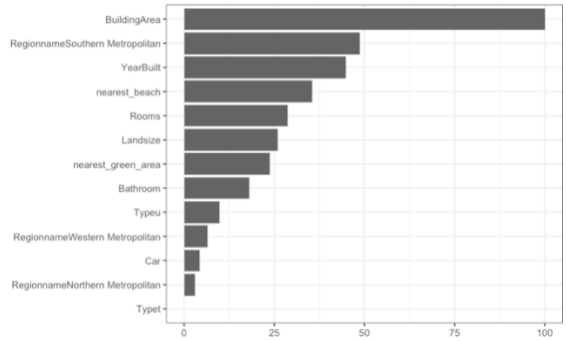

Figure 22: Random Forest Variable Importance

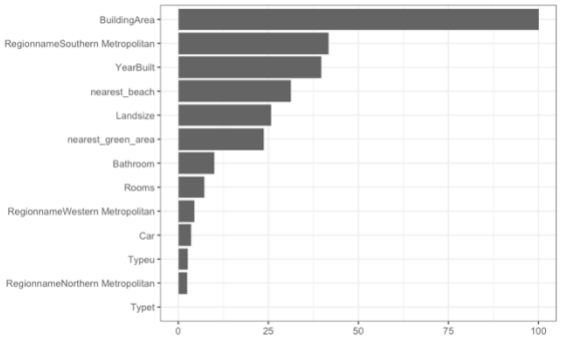

Figure 23: Gradient Boosting Variable Importance

As we can see, in both models, the area constructed is the most important variable, appearing in 100% of the trees constructed. Similarly, the following variables in level of importance are related to the region where the dwelling is located and its year of construction.

In the chaos of environmental variables, the distance to the nearest beach appears in about 30% of the trees in both models, while the distance to the nearest green area has a weight of slightly less than 25%.

Appendix J. Gradient Boosting PDP

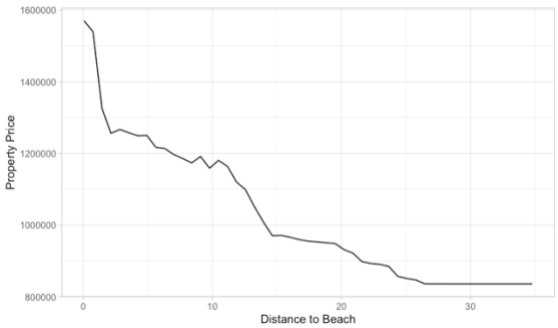

Figure 24: Boosting Distance to beach PDP

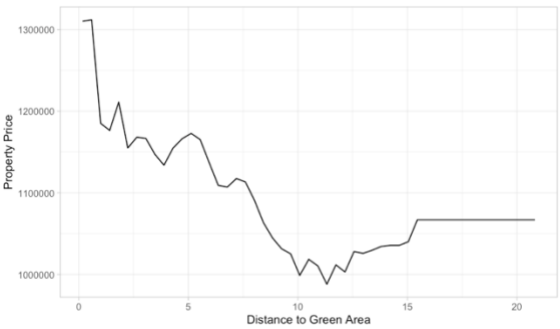

Figure 25: Boosting Distance to green area PDP