



Improving hedonic housing price models by integrating optimal accessibility indices into regression and random forest analyses

David Rey-Blanco ^{a,*¹}, José L. Zofío ^{b,c,1}, Julio González-Arias ^{a,1}

^a Department of Business Economics - UNED, Paseo Senda del Rey 11, 28040, Madrid, Spain

^b Department of Economics - UAM, Francisco Tomás y Valiente 5, 28049, Madrid, Spain

^c Erasmus Research Institute of Management - EUR, Burgemeester Oudlaan 50, 3062PA, Rotterdam, The Netherlands

ARTICLE INFO

Dataset link: <https://data.mendeley.com/datasets/3zzz8m5p3m/1>, <https://github.com/davidreyblanco/accessibility/tree/master>

JEL classification:

C14
C21
C55
R32
R14

Keywords:

Hedonic housing price models
Accessibility indices
Spatial dependence
Machine learning

ABSTRACT

Location indices are key in explaining variation in house prices. However, the definition of comprehensive indices capturing all locational features, along with their efficient and timely calculation, is usually one of the most complex dimensions of house price modeling. Existing difficulties result in partial location specifications, mostly due to three hurdles: (1) there is not a consensus on the best method to construct these indices, (2) what features (variables) to include: labor, demographic, commuting, etc., and (3) its creation requires granular and updated datasets. We introduce a methodology based on computer algorithms to create car and walk accessibility indices that address the previous concerns and capture location interactions among a wide range of variables. The selection of variables is based on an automated search of the best performing utility bearing gravitational accessibility indices for price prediction. Once these optimal indices are obtained, the method applies principal components analysis to secure their orthogonality. Using a unique dataset from a leading real estate portal in Europe, we illustrate and test for the city of Madrid their applicability in several house asking price models that are estimated using regression analysis and random forests (as a representative family of machine learning techniques). The experimental analysis reveals that using the optimal indices results in significant improvements in accuracy, for both regression-based models (13%) and random forests (21.6%), while achieving a substantial reduction in spatial autocorrelation (around 35%). The generated indices are clearly interpretable, which makes them a valuable tool for urban analyses (planning, transportation, sustainability, etc.). Finally, the methodology can be extended to other types of real state (commercial, industrial, etc.) and location (country, region, etc.).

1. Introduction

It is common knowledge that house location is key when individuals make the decision to buy a home, thereby determining its value. This intuition is backed by extensive research showing that households move mainly due to changes in job location (saving in commuting times) or changes in their family situation and income (requiring different housing characteristics)—for early references see Hanushek and Quigley (1979), Friedman and Weinberg (1981) and Malpezzi (2003). Hedonic price models (HPMs) allow quantifying the effect of location on house prices. In these models location is represented by a set of variables conforming the notion of *accessibility* to transport infrastructure, local public services, amenities, etc. Under traditional regression analysis, the parameters estimated through HPMs allow identifying the marginal contribution to price that each one of the considered attributes make.

However, there is general consensus that the misspecification of the accessibility variables included in (or excluded from) the model leads to a series of econometric problems resulting in biased estimations. Among the foremost problems we highlight spatial heterogeneity and spatial autocorrelation (Anselin & Griffith, 1988), multicollinearity (Orford, 2017), and heteroscedasticity (Fletcher, Gallimore, & Mangan, 2000).

All these problems, particularly spatial heterogeneity and spatial autocorrelation, derive from a poor specification of location attributes in the HPM. In a systematic review of how location is incorporated in hedonic models, Heyman, Law, and Pont (2018) stress that the majority of models rely on poorly elaborated location variables or arbitrarily aggregated area features. In our view this is the result of three factors: data unavailability, inappropriate spatial partitioning and computational complexity of generation. The bottom line is that these variables are often incomplete, outdated or arbitrarily specified.

* Corresponding author.

E-mail addresses: drey7@alumno.uned.es (D. Rey-Blanco), jose.zofio@uam.es, jzofio@rsm.nl (J.L. Zofío), jglez@cee.uned.es (J. González-Arias).

¹ In memoriam of Juan Antonio Vicente-Vírseda, whose scholarship and commitment to research and education will be greatly missed.

When addressing the question of how to define proper accessibility indices, the first challenge to address is how to specify the location variables, bearing in mind that urban layouts are diverse and the influence of factors vastly vary from one area to another. The second question is the availability of information: data cannot be timely collected and may be aggregated with arbitrary criteria for the domain of the problem. For instance, census tracts are designed for describing population distribution and characteristics, but not for capturing features of their housing submarkets. The third concern is that simplistic definitions of accessibility indices to location attributes do not reflect the marginal utility of location, as they are normally based on ‘as the crow-flies’ Euclidean distances or relatively better travel times. However, satisfactory indices reflecting the utility of the location based on gravitational specifications are rarely used in the industry mainly due to the computational cost of calculating them.

In this study we propose and evaluate a methodology for constructing a set of gravity-based accessibility indices for HPMs that overcome existing limitations. These indices address the main issues met when defining this kind of variables; namely, making the creation process *systematic*, being consistent with a *utility bearing framework*, and generating *spatially robust* variables. Each location accessibility index summarizes numerically the opportunities that a real state property has nearby, such as access to workplaces, transportation, leisure, schools, etc. To improve their statistical properties we generate a set of orthogonal accessibility indices. The method relies on the principal components analyses to produce these indices, ensuring that they control for spatial dependence, as they are later used as explanatory variables in the HPM. We illustrate their applicability for predicting house prices in Madrid (Spain), and check their performance in terms of accuracy and predictive power using a *naïve* benchmark model that is compared to two increasingly complex models in terms of the inclusion of locational attributes (either through dummies as in standard specifications, or through our newly proposed accessibility indices). Using common metrics, e.g., mean average percentage error, our results confirm the superiority of the automatically generated accessibility indices when predicting housing prices. We perform several robustness checks by using four different estimation methods, two related to regression analysis and two based on trees and random forests (as representative families of machine learning algorithms).

Machine learning techniques are becoming increasingly common in fields where traditional techniques such as statistics were exclusively applied. One main reason for this shift is that, from a predictive perspective, deep learning offers excellent levels of accuracy for complex matters (Han et al., 2022; Huang, Zhang, Wang, Wu, & Song, 2022; Huang, Zhang, Wu, Min, & Song, 2022; Kumar & Suresh, 2023; Tang, Zhang, Min, & He, 2022). These methods are especially useful to study real estate markets when applied to large volumes of data from classified advertisement portals. For example, Bricongne, Meunier, and Pouget (2023) track daily prices on data from five portals in the United Kingdom using machine learning techniques. In Asia, Wang, Li, and Wu (2020) develops a housing price index for 274 Chinese cities based on data from real estate websites. Therefore, although regression analysis represents the most common estimation method, machine learning techniques have been gaining ground in the prediction of housing prices since the nineties. First approaches were performed with neural networks (Kauko, Hooimeijer, & Hakfoort, 2002; Liu, Zhang, & Wu, 2006; McCluskey & Anand, 1999; Pace, 1995; Selim, 2009). However, there is a growing number of studies based on tree-based algorithms, especially those based on ensemble models, such as gradient boosting or random forests. Tree models have demonstrated their suitability to the problem (Antipov & Pokryshevskaya, 2012), with a good performance compared to other techniques (Steurer, Hill, & Pfeifer, 2021; Valier, 2020) and making it possible to perform accurate massive appraisals with a large degree of explainability. Specifically, random forests (hereafter RF) have been widely used for modeling house prices, either on academia or industry. Among them, Antipov

and Pokryshevskaya (2012) benchmark a RF model against ten other machine learning algorithms, over properties on sale in Saint Petersburg (Russia). Baldominos et al. (2018), Alfaro-Navarro et al. (2020) and Rico-Juan and Taltavull (2021) use RF models to build house price models for Spanish properties. As aforementioned, we contribute also to this literature by comparing the performance of regression methods like the Lasso and Elastic-Net Regularized Generalized Linear Models, LERG, against recursive partitioning trees and RF. In the comparison we include the optimal accessibility indices obtained by applying two algorithms aimed at maximizing their explanatory power, while removing collinearity among them. Our results indicate that the regression models equipped with these optimal indices perform as well as the machine learning methods, suggesting that the new methods are capable of generating indices that properly capture the influence of locational attributes on price. We confirm this by establishing the spatial robustness of the results, which is also on par for both models. This is in sharp contrast to previous findings where machine learning clearly outperforms classical regression, but where the use of optimal accessibility indicators is not incorporated to the modeling process. Hong, Choi, and Kim (2020) shows a fourfold absolute error reduction in automatic valuations in the Gagnam neighborhood in Seoul (South Korea), while Čeh, Kilibarda, Lisec, and Bajat (2018) and Hjort, Pensar, Scheel, and Sommervoll (2022) report similar superiority.

The three main methodological contributions of our research are, firstly, the proposed method’s approach to address *ex-ante*, and from a utility perspective, the bias introduced by spatial dependence. This contrasts with standard spatial analysis approaches whose primary objective is to deal *ex-post* with its existence and effect on a studied phenomenon. As explained by Montero, Fernández-Avilés, and Mateu (2015): “...in geostatistics, the most important aspect in the geostatistical analysis is to quantify the spatial correlation between observations (...) and use this information to achieve the previous objectives...”. Thus, improving existing calculations of accessibility indices, our contribution has the advantage of being able to better capture the effects of locational attributes on housing price models by filtering out spatial dependence. In this regard, our methodology does not resort to ad-hoc spatial statistical measures but addresses its root causes (Pot, van Wee, & Tilrema, 2021), introducing a valuable econometric analysis tool. Secondly, the methodology is implemented through efficient algorithms: as the orthogonal (or boosted) accessibility indices are precalculated over a fixed grid, our method puts forward an inexpensive computational process to bring detailed location features in house valuation problems, making it suitable for highly demanding data setups on any database engine. Thirdly, we show how embedding our newly proposed accessibility indices into a standard hedonic price model improves the results in terms of accuracy and spatial robustness, regardless the estimation approach, either using regression methods or more recently available machine learning techniques.

The article is structured as follows, after this introduction we review the literature on the use of location attributes in hedonic housing price models; in particular the use of accessibility indices. The third and fourth sections describe the methodology and data sources for calculating the newly proposed optimal accessibility indices. In the fifth section we estimate a series of hedonic price models through econometric methods and machine learning techniques, and compare the results obtained with both approaches. Here we show that the inclusion of the new accessibility indices improves the performance of regression methods when compared to trees and RF. In the sixth section we further reinforce the advantages of the proposed methodology by looking into how capable they are at minimizing spatial autocorrelation. We draw the main conclusions and identify further lines of research in the last section.

2. Literature review on locational attributes for hedonic housing price models

In this section, we carry out a literature review on the topic of modeling the effect of locational attributes on hedonic price modeling through accessibility indices. We survey classic references and newly proposed approaches, both from the conceptual and empirical perspectives. In particular, the emergence of automated processes and machine learning techniques in the calculation of the accessibility indices entering the estimation of hedonic prices.

Although accessibility has been a central topic on physical urban planning from the second half of 20th century, first uses of this term date back to the 1920s. However, it was Hansen (1959) who proposed a preliminary methodology for the use of accessibility for urban planning. Since then a location's accessibility refers to the *intensity of potential interactions* with a series of opportunities or attributes such as employment, public services and amenities. Related studies in other fields as population geography (Stewart, 1947) first defined the gravitational potential of a location by weighting a sum of forces related to the rules of distribution and population equilibrium. Accessibility indices associate a given opportunity at a specific place with the cost of realizing it (Batty, 2009). This cost, also called impedance, is normally measured as distance or, at best, travel time. Accessibility indices are usually presented in an aggregate or composite form that summarizes how easy or difficult it is to realize a given level of utility, represented by a vector of opportunities at the place of interest. However, due to data constraints and computational limitations, the definition of these indices is normally done in a simplistic way without resorting to gravitational models. Our methodology develops these methods by proposing the definition of optimal accessibility indices based on a utility-bearing gravitational approach that is computationally implemented in an efficient manner through automated algorithms.

Even if some authors argue that the adoption of accessibility in urban planning has not evolved much in the past decades, e.g. Handy (2020), some promising studies reintroduce the accessibility concept as the framework of thinking in urban analysis. In this regard, the increasing abundance of open sources and new computational capabilities have brought up new generic spatial features (Vecchio & Martens, 2021). Examples of the latter are, for example, the set of global generic indicators proposed by Boeing et al. (2022), or the urban spatial features called 'Spatial Signatures' put forward by Arribas-Bel and Fleischmann (2022). All these sources have been also released publicly for further analyses (Samardzhiev, Fleischmann, Arribas-Bel, Calafiore, & Rowe, 2022).

A hedonic price model acknowledges that a heterogeneous good can be described by a series of attributes. Therefore a good is essentially a set of characteristics whose value can be ascertained by the aggregated performance of utility-bearing characteristics (Rosen, 1974). In the case of real estate property such as housing, location results in a series of advantages or disadvantages generating utility or disutility, which affects properties' sale price. Consequently, when specifying hedonic price models for housing, it is imperative to include accessibility indices. The first approaches to introduce locational attributes as part of hedonic models date back to Kain and Quigley (1970), who included a set of neighborhood characteristics including distance to the central business district (capturing the land-rent gradient by which the farther are locations from the CBD, the lower are their prices due to higher transportation costs), along with the structural characteristics of the housing unit. Later on these models were enriched with other features (see Bowen, Mikelbank, & Prestegard, 2001) by including market variables related to supply and demand characteristics (number of properties on sale, intensity of demand, etc.). There is consensus in the literature that simple popular methods for encoding locational attributes like using dummies have the implication of limited explanatory power compared with relative location variables, as shown by Heyman and Sommervoll

(2019) when explaining housing prices in Oslo, Norway. Our study further investigates this question by comparing the results obtained for several models that differ in the treatment of locational attributes: excluding them, using simple dummies, or including optimal accessibility indices. Our contribution helps overcoming the imprecise specification of location attributes in standard hedonic modeling by introducing interpretable, highly-granular, easy-to-calculate indicators capable of producing better-fitted regressions (Diewert & Shimizu, 2021).

Lately a great number of studies have found empirical evidence of the relation of locational attributes and house prices from diverse standpoints. To name a few, Cao (2015) showed that closeness to industrial, commercial, multifamily and public land uses tend to increase surrounding home values. Hence, this author concludes that a balanced (optimal) mix of land use activities should be sought when locating economic activities into neighborhoods. Also, accessibility to public services and transportation have received special attention. Access and quality to public services have been found to be significant determinants of housing value (see Lan, Wu, Zhou, & Da, 2018). This research finds relevant the access to municipal services, average academic results and student-teacher ratios. Bowes and Ihlanfeldt (2001), Bartholomew and Ewing (2011), Agostini and Palmucci (2017), and Choi, Park, and Uribe (2022) study the effect of access to public transport, including both direct and indirect effects of transit stations, on the attractiveness of nearby neighborhoods. They found that stations located away from downtown have positive impacts, creating 'islands' of higher property values. In the same vein, Zhou, Chen, Hong, and Zhang (2021) analyze the effect on house prices of introducing a new subway station in Shanghai. By reducing travel time, easier commutes to the central business district result in an average house price increase of 3.75%, with the most distant residential zone enjoying the largest price growth. Recently Li (2020), using several data sets of Beijing's congestion patterns and housing prices, find that consumers are willing to pay significantly more for access to rail transit in more congested areas. He corroborates the prediction that the expansion of the metro network mitigates the costs of road congestion, creating both private and social benefits. Our contribution, in this regard, tackles the difficulty of selecting the best definition for public transport attributes to include in price models, by using the proposed automated framework.

Besides the different attributes surrounding the property determining its value (in terms of utility for the buyer), it has been also shown that the street layout may have a relevant effect on house prices. This is particularly relevant for differentiating accessibility in terms of driving times or walking times. Asabere and Harvey (1985) provided empirical evidence from Halifax-Dartmouth, Canada, concluding that an open street layout, such as a residential zoning with large lots and high road density, leads to higher prices. Studying residential property valuation in Minnesota, Iacono and Levinson (2011) give evidence about the positive effect that a lower distance to major highway links (access points) has on house prices. Datagov (2019) shows that pedestrian walkability is related to greater housing prices, by capturing how the density or concentration of attributes is higher when they are within walking distance, calculated as the percentage of households less than a quarter of an hour walk to commercial use or a public transport stop. The contribution of 'walkability' to house prices might be different depending the specific geography of the urban areas, which is taken into consideration through Geographical Information System (GIS) techniques—early studies in the US carried out by Yates and Miller (2011), who rely on a Walk ScoreTM, and Sohn, Moudon, and Lee (2012), confirmed that a combination of pedestrian infrastructure and land use mix significantly contributes to increases in rental multi-family property values. Recently, Choi et al. (2022) reinforce these results, finding that walkable neighborhood designs, when coupled with (light) rail accessibility, have significant and positive impact on values for all residential properties. Few studies make the distinction between car and pedestrian accessibility, choosing just one of them. Since considering both types of accessibility is relevant when studying

house prices in urban areas we also differentiate between the two in our methodology and empirical application. Also, in contrast to ad-hoc approaches, our modular algorithm design facilitates the introduction of future new transport modes or spatial features.

Finally, Heyman et al. (2018) perform a meta-analysis of the literature, finding that even if a majority of hedonic housing price models include accessibility variables in their specifications, they do so in a simplistic way. In particular, the preferred use of simple Euclidean distance or even travel times, given their calculation ease, over gravity-based indices, questions the ability of the models to capture the real effect of location on prices. Also, although Euclidean distance to CBD is a straightforward method to introduce accessibility, this monocentric approach seems no longer to be valid for current urban configurations as we find in the studies by Waddell, Berry, and Hoch (1993) and Xiao (2017). These authors challenge the validity of monocentric schemes suggesting the use of polycentric models. In order to address the question, Small and Song (1994) suggested a number of accessibility, demographic and urban planning measures to be included in a housing price model. Later work by Knaap and Song (2003) applied a series of quantitative measures using GIS to capture the contribution of different features to price. They defined six characteristics affecting single-family homes: street design and circulation systems, connectivity, block size and square mesh configuration. We consider all this received knowledge in our proposal of optimal accessibility indices by considering the most advance GIS techniques when including travel time as impedance variable in the gravity-based specifications—both by car and walking.

3. A new methodology to create automated optimal accessibility indices for hedonic price modeling

In this section, we present the methodology, based on a utility-bearing gravitational approach, to calculate a family of raw accessibility indices for each locational attribute, which are defined for different impedance values. How to choose among the different possibilities to determine the optimal specification of these indices is presented in the following section. However, since these indices are intended to be used as explanatory variables in hedonic housing price models, we start presenting their standard specification and how accessibility indices enter its formulation. Here we also discuss the statistical sources, datasets, and geographical information systems employed in the calculation of the accessibility indices.

3.1. Specification of the hedonic housing price model

Our proposed methodology creates a set of accessibility indices that capture the contribution of location to house prices. In the next section we present the unique dataset we use to create these indices. The dataset is compiled in collaboration with *Idealista*, a leading real estate portal in Europe. We resort to a standard hedonic housing price model to illustrate and test the relevance of the new indices, and decompose the contribution that different attributes make to the price of real estate property. As we show below the model is estimated using both econometric and machine learning methods, to check the robustness of our results regarding the relevance of the novel accessibility indices. The specification of these indices follows the categorizations proposed by Heyman et al. (2018). Hence, an accessibility index is defined by a series of opportunities a given location offers and an impedance (gravitational) function based on the transportation cost from the opportunities to the location, which is measured in travel times.

Economic theory provides guidance regarding the right functional form of the hedonic housing price models (Freeman, 1979; Rosen, 1974). Cassel and Mendelsohn (1985) discuss the pros and cons of the Box–Cox specification, which generalizes familiar forms like the semilog, log linear or translog. However, they stress that the large number of parameters involved in this specification reduces the accuracy

of the estimates, it is not suited for negative data, and maybe inappropriate for prediction purposes. Since our main goal is to determine the eventual superiority of our proposed accessibility indices in terms of accuracy and predictive power over simplistic options, keeping the model specification as simple as possible serves our purpose. Abiding by the principle of parsimony in this dimension of the study, we rely on the standard formulation represented by the *time dummy hedonic model*—see, e.g., Eurostat (2012, Ch.5).

This formulation expresses the model as a linear function of the property attributes, which based on the rich available information corresponds to four categories of characteristics—Table 1. The hedonic house price model assumes the price of a property n in period t , p_n^t , is a function of a fixed number of characteristics or features $q = 1, \dots, Q$, reachable by car or walking, $m=\{\text{car, walk}\}$, which are measured by a series of quantities $Z_{nq}^{tm} = \{\hat{A}_{nk}^{tm*}, S_{nj}^t, M_{nl}^t, D_{nk}^t\}$ observed at $t = 1, \dots, T$ periods, plus a random error term ϵ_n^t . That is,

$$p_n^t = \beta_0 + \sum_k \beta_k \cdot \hat{A}_{nk}^{tm*} + \sum_j \beta_j \cdot S_{nj}^t + \sum_l \beta_l \cdot M_{nl}^t + \sum_t \delta \cdot D_n^t + \epsilon_n^t, m = \{\text{car, walk}\}, \quad (1)$$

where: (1) \hat{A}_{nk}^{tm*} represents our locational *accessibility (A)* indices in terms of $k = 1, \dots, K$ locational *opportunities* that are reachable by car and walking; (2) S_{nj}^t denotes the $j = 1, \dots, J$ *structural (S)* attributes of the property; (3) M_{nl}^t captures the $l = 1, \dots, L$ supply and demand features of the submarket (*M*) to which the asset belongs; and (4) D_n^t accounts for the *time dummies (D)*. We anticipate here that the \hat{A}_{nk}^{tm*} accessibility indices entering model (1) are the result of an optimization process that, departing from a set of *raw accessibility* indices, A_{nk}^{tm} (i.e., without the ‘hat’ notation), determines their best definition by maximizing their correlation with the residuals of a *naïve OLS* model that omits accessibility but includes the rest of attributes ($S_{nj}^t, M_{nl}^t, D_{nk}^t$), and subsequently transforms these *correlated* indices, A_{nk}^{tm*} (termed ‘optimal’ and denoted with the superscript ‘*’), into *orthogonal* ones through principal components analysis—ultimately obtaining \hat{A}_{nk}^{tm*} , which constitute what we call *boosted accessibility* indices. Also, besides the automated generation of these orthogonal indices \hat{A}_{nk}^{tm*} to be included in (1), a further contribution of this study is the introduction of the market variables M_{nl}^t . These are rarely available in empirical modeling due to the difficulty of obtaining this information, however they are instrumental in understanding the behavior of supply and demand forces for each (sub)market (Piazzesi, Schneider, & Stroebel, 2015). As shown in the next section, we are capable of identifying and quantifying many of these characteristics. Finally, the dependent variable p_n^t is price per square meter, rather than overall price, as it helps reduce heteroscedasticity in the model. Under classic error assumptions, in particular a zero mean and constant variance, the model is estimated on the pooled data pertaining to all time periods, represented by the dummies.

3.2. Specifying and calculating the raw accessibility indices A_{nk}^{tm}

We now focus on the definition of the variables entering the accessibility indices from the input sources that numerically summarize the opportunity set that a location brings to individuals. As anticipated, in densely populated areas, it is relevant to consider accessibility by car and foot (Wee & Vickerman, 2021). The construction of each pair of drive and walk indices is based on the same attributes for both transport modes and, as shown in the next section, we let the optimizing algorithm decide their importance for each mode. In this way we prevent the introduction of subjective biases that would entail the choice of different attributes for each index. Each opportunity entering the accessibility index is constructed as a gravitational index of the values of each variable within a series of isochrones, which are adapted according with the two transport modes, $m = \{\text{car, walk}\}$. For the car transport mode we consider the vector of distances $d_{i(m=\text{car})}$, reachable

Table 1
Hedonic model variable categories.

Category	Motivation for inclusion
Locational accessibility: \hat{A}_{nk}^{tm*}	Optimal accessibility indices obtained by applying principal components analysis to a set of optimized raw accessibility indices. These indices capture the contribution of location to the price of a property, including accessibility indices based on neighborhood characteristics and other geographical features
Structural characteristics: S'_{nj}	Structural variables control for the physical features of the property such as square meters, room number or state of conservation
Market features: M'_{nl}	Incorporates the main supply and demand dynamics of the market where the property is located
Time dummies: D'_n	measure adjustments of price over time, seasonality and trend effects

from every location in the following times: $i(\text{car}) = \{1, \dots, I\} = \{5, 10, 20, 30, 40, 50, \text{and } 60 \text{ minutes}\}$, whereas for the pedestrian indices we consider $d_{i(m=\text{walk})}$ with the following times: $i(\text{walk}) = \{1, \dots, I\} = \{5, 10, 20, \text{and } 30 \text{ minutes}\}$. Driving distances are calculate assuming the legal driving speed limit, while walking distances assume a speed of 5 km/h.²

When defining the raw accessibility indices A_{nk}^{tm} our conceptual framework assumes that locations yield utility to home owners through access to a set of opportunities represented by a set of specific variables observed in location n at time t . The utility yielded by an opportunity is decreasing in transportation costs, which we model through an exponential penalty function which is inversely proportional to the distances reachable by the alternative travel times. Specifically, for each variable k , our raw index of location accessibility, A_{nk}^{tm} , aggregates its values for all I isochrones (e.g., number of bus stops within 5, 10, 20 and 30 min walking times), each weighted by its relative travel time, into a single scalar. This corresponds to the following specification (see, e.g., Levinson & Krizek, 2005):

$$A_{nk}^{tm} = \sum_{i(m)}^I Opportunity_k(X, Y, d_{i(m)}) \cdot e^{-\beta_m \cdot d_{i(m)}}, m = \text{car, walk}, \quad (2)$$

where $Opportunity_k(X, Y, d_{i(m)})$ represents a specific opportunity (variable) located at coordinates X, Y and distance d within the geographical limits of the isochrones associated to the location. The distance measure $d_{i(m)}$ corresponds to the aforementioned ranges of travel times (in minutes). The parameter β_m represents the exponential decay of the index for the applied impedance (in this case the calculated travel time distance). The accessibility index (2) is specific for each transportation mode, either walking or driving. In the empirical section we discuss the selection of the optimal values of A_{nk}^{tm*} for a range of β_m values entering the optimizing algorithm.

3.2.1. Opportunities and specific variables entering the location accessibility indices

According to Eq. (2), our gravitational accessibility indices A_{nk}^{tm} are constructed by aggregating the number of items for each $k = 1, \dots, K$ opportunity observed in a given location n at time t . We present in Table 2 the $K = 26$ variables used in this study to represent these opportunities. These variables are grouped into five categories: *Public transportation*, *Private transportation*, *Economic activity & Basic services*, *Social and Recreational*. We define $Opportunity_k(X, Y, d_{i(m)})$ as generally as possible, and therefore adopt the classification set out by Heyman et al. (2018). This classification includes the most commonly used variables considered by the industry and academia. Therefore, an

$Opportunity_k(X, Y, d_{i(m)})$ index is in essence a measure of a variable contained within an isochrone at time-distance $d_{i(m)}$, whose measurement is performed cumulatively at one of the following levels: as a count of elements, as a sum of areas, as a sum of lengths or as a density, specified as:

$$\begin{aligned} Opportunity_k(X, Y, d_{i(m)}) &= \sum_{o=1}^O M(X, Y, o_x, o_y, d_{i(m)}) \cdot C_k(o_x, o_y), \forall o \in O, \\ M(X, Y, o_x, o_y, d_{i(m)}) &= \begin{cases} 1, & \text{if } distance(X, Y) \rightarrow (o_x, o_y) \leq d_{i(m)} \\ 0, & \text{otherwise} \end{cases} \end{aligned} \quad (3)$$

where for each item o in the complete opportunity universe O , with a pair of coordinates (o_x, o_y) , we define a contribution measure C_k , whose definition depends on the aggregating function to apply on each k family (count, sum, or density). $M(X, Y, o_x, o_y, d_{i(m)})$ is a dichotomic function used to filter all eligible contributing opportunities at a distance $d_{i(m)}$.

We briefly comment on the different groups of location opportunity indices (2). A detailed definition is presented in on-line Appendix A.

- *Public and Private transportation indices.* The opportunity accessibility indices for public transport aggregate the elements of each kind (bus stops, metro stations, etc.) at a certain impedance time-distance. Private transportation opportunities use the length in meters of highways and the density of the road grid in square meters. These latter measures result in increased utility, by providing better connectivity in the suburbs of a city, or as a disutility as it involves higher levels of pollution and noise.
- *Economic activity & Basic services indices.* These opportunity accessibility indices refer to access to economic activities, basic services, residential facilities, employment and leisure. It is noticeable the presence of many of these variables in particular locations. For instance, hotel, food, tourism and vacation rentals are usually found in high numbers in specific touristic areas. Considering low granularity indices by way of increasing isochrones is important to prevent geographical bias resulting from the modifiable areal unit problem (MAUP). In this regard some measures are based upon the sum of total area of certain land uses. The rationale behind this choice is the assumption that the number of square meters of residential properties would act as a proxy of population, whereas the number of meters of office, industrial and commercial space are a proxy of gross employment concentrations or labor subcenters (Giuliano & Small, 1991).
- *Social and Recreational indices.* These opportunity accessibility indices capture relevant socioeconomic variables and recreational amenities, respectively.

3.2.2. Statistical sources and dataset

We rely on six statistical sources to construct the range of $Opportunity_{i(k)}(X, Y, d_{i(m)})$ accessibility indices entering the hedonic housing price models. Data refers to the metropolitan area of Madrid (Spain). With 6.6 million inhabitants in 2020 (829.84 inhabitants/km²),

² The definition of isochrones ranging from 5- to 30-minutes on foot, or from 5- to 60-minutes driving, are commonly used in the accessibility literature. For example, Ewing and Cervero (2010) and Handy and Niemeier (1997) suggest using 10-minute walking thresholds to measure accessibility to points of interest under usual urban configurations, while Frank et al. (2010) recommend taking 5-minute walk distances to parks and transit stops, and longer boundaries for other destinations.

Table 2
Catalog of raw opportunity indices.

Category	Opportunity: $Oportunity_{i(k)}(X, Y, d_{i(m)})$	Variable: k	Measure	Source	Car	Walk
Public transportation	bus	TRANSPORT.BUS	count	OSM	x	x
	metro	TRANSPORT.METRO	count	OSM	x	x
	railway	TRANSPORT.TRAIN	count	OSM	x	x
	airport	TRANSPORT.AIRPORT	count	OSM	x	
Private transportation	highway	ROUTING.HIGHWAY	length	OSM	x	x
	routing	ROUTING.COMPLEXITY	density	OSM	x	x
Economic activity & Basic services	land	CAD.URBANLAND	area	cadastral	x	x
	hotel	HOTEL	count	OSM	x	x
	hotel	VACATION	count	airbnb	x	x
	food	FOOD	count	OSM	x	x
	tourism	TOURISM	count	OSM	x	x
	education	CAD.PUBLIC	area	cadastral	x	x
	education	CAD.SCHOOL	area	cadastral	x	x
	education	EDUCATION	count	OSM	x	x
	tourism	TOURISM	count	OSM	x	x
	health	CAD.HEALTH	area	cadastral	x	x
	commerce	SHOP	count	OSM	x	x
	commerce	CAD.COMMERCE	area	cadastral	x	x
	agriculture	CAD.AGRICULTURE	area	cadastral	x	x
	venues	CAD.VENUES	area	cadastral	x	x
Social	religion	CAD.RELIGION	area	cadastral	x	x
	residential	CAD.RESIDENTIAL	area	cadastral	x	x
Recreational	park	PARK	count	OSM	x	x
	sport	CAD.SPORT	area	cadastral	x	x
	sport	SPORT	count	OSM	x	x

Table 3
Summary of data sources.

Source	Description	Record count
Cadastral	Number of parcels of any land use	569,791
	Number of residential parcels	482,203
	Number of residential properties	2,878,748
Open Streetmaps	Points of Interest (POIs)	38,141
Airbnb	Total listed properties	31,142
Isochrones	Polygons	344,069
	Seed locations	63,187
Idealista.com	Total listings on Idealista.com during 2018	1,781,568
INE	Census tracts	4,272
	Districts	246
	Municipalities	179

Madrid is the most dense populated area in the center of Spain ([Eurostat, 2020](#)). It also exhibits one of the most active real estate markets in Europe. Our dataset covers the major aspects of the housing market: properties on market, land uses, neighborhood characteristics, transport network and touristic offering. [Table 3](#) summarizes the main information on the statistical sources.

Idealista data constitutes also a relevant source of rich data at the individual property level. It gathers monthly information from a large set of ads of residential properties on sale in 2018. The site <https://www.idealista.com/> is the major listing of real estate in Spain with a leading market share. Cadastral data is disaggregated at parcel and property levels ([Registro Central del Catastro, 2018](#)). We use residential and non residential records. Each cadastral record contains: construction quality, age and physical and structural features. When it comes to socio-demographics and educational levels, we process public data records from the Spanish national institute of statistics ([INE, 2018](#)—on-line [Appendix B] lists the variables included in the idealista dataset. Open Street Maps supplies key information on the network transportation graph and points of interest ([OpenStreetMap](#)

[contributors, 2017](#)).³ The former is used to generate the time-distance isochrones for the two transport means of choice: car and walk. Finally, we use [Airbnb \(2017\)](#) data on vacation rentals to incorporate the effect of short-term rental market in the model. Cadastral data makes also another important contribution to our study as it helps identify all *seed locations* which are taken as reference for the calculation of the accessibility indices. Seed locations are based on the placement of residential parcels as discussed in Section 3.2.4 below.

3.2.3. Idealista listing data

The variables taken from the Idealista dataset comprise all active ads of residential use (single and multiunit homes) in the Madrid region. [Table 4](#) details all the relevant variables corresponding to the Structural S'_{nj} and Market M'_{nl} dimensions included in the hedonic price model (1) (the Structural category comprises the variables referring to *Location, Unit, Building* and *Quality*).

³ Open geographical databases are increasingly used in hedonic models. For instance, [Xiao \(2017\)](#) uses OSM POI for house a price model for Beijing (China) to tackle spatial autocorrelation.

Table 4
Idealista dataset description.

Class	Variable name	Description
Date	PERIOD	Ordinal month code from 1 to 12, mapped to the month in 2018 when the observation was taken
Location	LOCATIONID	Idealista website area unit equivalent to administrative divisions
	LONGITUDE	Longitude in CRS EPSG:4326
	LATITUDE	Latitude in CRS EPSG:4326
Unit	CONSTRUCTEDAREA	Total area in square meters
	ROOMNUMBER	Number of rooms
	ISSTUDIO	is it a studio apartment?
	ISDUPLEX	it is a two-story flat?
	HASANNEX	has it a boxroom or parking space?
	HASAIRCONDITIONING	has airconditionining?
Building	FLOOR_POSITION	Vertical location of the property in building: B lower stories, M middle part or T when located in the upper part
	HASLIFT	has it an elevator?
	HASSWIMMINGPOOL	has its building a swimming pool?
	CONMSTRUCTIONYEAR	Year of construction
Quality	MAXBUILDINGFLOOR	Number of floors in building
	CADASTRALQUALITYID	Construction quality, numerical from the best 1 to the worst 9
Market	BUILTYPEID	Condition: new, second hand not renewed, second hand renewed
	PRICE	Price in euros
	CHANNELID	Sales channel for the property: 1 Real Estate agent, 2 bank-owned property and 3 individual seller.
	UNITPRICE	Price in euros by square meter
	ONMARKET_RENT	Properties on sale on Idealista/Total built residential properties in LOCATIONID
	ONMARKET_SALE	Properties for rent on Idealista/Total built residential properties in LOCATIONID
	RENTSALE_RATIO	Properties for rent/on sale on Idealista website
	DEMAND	Demand intensity, derived from the number of monthly web views of the advert

To prevent the detrimental effects of anomalous values or repeated advertisements we performed an outlier cleanse and deduplication process. It is usual finding multiple ads of the same property since it is a common practice to sell them through several agents. On the other hand we transform all categorical variables to dummy variables; e.g. in the case of time the *PERIOD* consists of 12 monthly dummy variables for 2018 (see on-line Appendix B).

3.2.4. Discrete space and flexible granularity

We now describe further qualifications of the GIS methodology aimed at addressing the course of dimensionality, which is one key issue in spatial analysis. When estimating the hedonic price model (1) the dimensionality problem emerges from the large number of combinations of elements that interact within a designated area; in our case the many $Opportunity_k(X, Y, d_{i(m)})$ accessibility features corresponding to the number of public transportation stops (bus, tram, train), economic activities and public services (commerce, education, health), etc. By reducing the dimensionality of the analysis we can keep the computation time of the algorithms obtaining the optimal accessibility indices within reasonable bounds. This is achieved by limiting the number of areas considered, thereby restricting the location seeds included in the geographical grid. A criterion compatible with this reduced dimensionality (which nevertheless is adopted in many studies) is to set as location seeds those corresponding to the coordinates of all the centroids of cadastral parcels with residential dwellings (i.e., industrial areas are excluded since we focus on residential property). These administrative centroids constitute the precise locations from where we calculate the different isochrones and their associated accessibility measures.

We reduce the number of seeds without compromising the required geographical precision. For this goal we consider two granularity levels for the location seeds depending on the transportation mode. Specifically we consider a Geohash H3-10 level resolution for walking and a 9 level resolution for driving (see <https://h3geo.org/>). Areas with this

resolution are enough for the analysis. In the case of H3-10 resolution the average error in distance to a specific landmark is 50 m, corresponding on average to half a minute walking (a minute by foot at the standard pace covers 100 m), and far less than that by car. In summary, as shorter than a minute travel time does not make a difference for individuals, we believe the chosen resolutions are adequate for walking and driving, which is compatible with the limitation of the number of areas considered to address the dimensionality issue. That way we find a suitable balance by trading unnecessary geographical granularity for gains in computational performance. Fig. 1 illustrates the generation of cadastral location seeds for the whole Madrid region.

Following existing methods, the geographical information on the attributes is translated from the continuous coordinate space (generally represented as a vector of real numbers: latitude and longitude) into a discrete space. Our approach relies on an existing Discrete Global Grid System *DGGS*, that employs a tessellated spaced composed by polygonal (hexagonal) shapes (Bondaruk, 2019). The *DGGS* creates a numerical identifier for each hexagon. Also, for technical convenience we use Uber H3 tessellation (Uber, 2018) as it is available in a number of databases and programming languages.

Fig. 2 illustrates the degree of granularity for accessibility measures considering the 9 level resolution for driving times (left) and the 10 level resolution for walking times (right). For the 9 level resolution hexagons have an edge length of approximately 174 m, while in the 10 level resolution they have 66 m.

Finally, we adopt an additional criterion to further reduce the dimensionality problem by assuming that the influence of differences in transportation cost *within* the isochrone crown defined by the 5 and 10 min walking are not relevant for the analysis. This implies that individuals do not perceive a substantial change in the utility level if we take different destination points in this range. The series of raw accessibility indices used in this study, corresponding to the variables presented in Table 2 and calculated according to Eq. (2) for the city of Madrid using the H3 grid at 9 and 10 resolutions, are available at <https://github.com/davidreyblanco/accessibility/tree/master>.

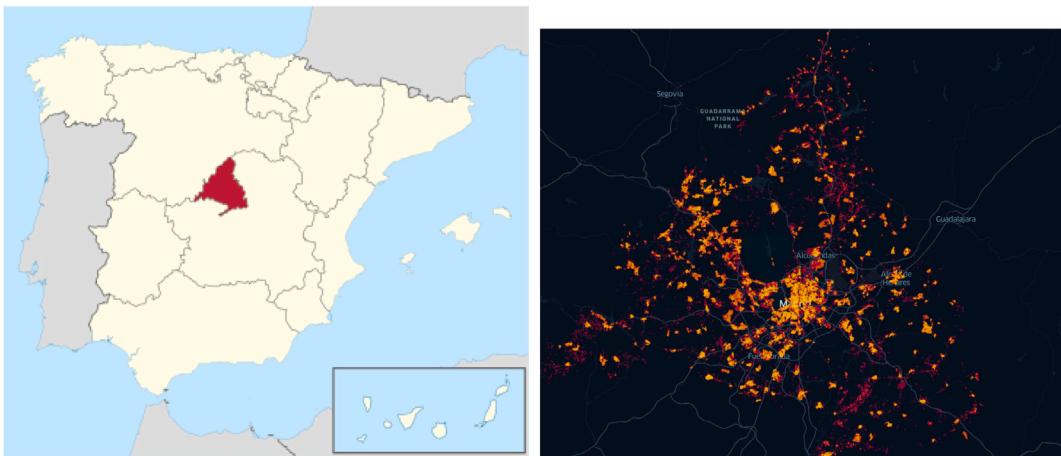


Fig. 1. Comunidad de Madrid (left) and cadastral seed locations (right).

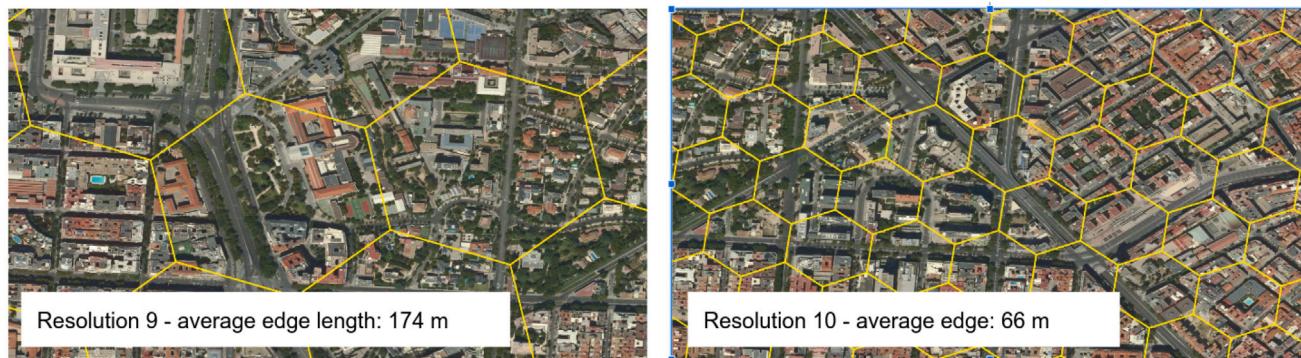


Fig. 2. H3 resolutions used for the seed locations. We used resolution 9 reachable opportunities by car and level 10 reachable on foot.

4. Calculating the optimal (or best of breed, A_{nk}^{tm*}) and orthogonal (or boosted, \hat{A}_{nk}^{tm*}) accessibility indices: Algorithms

In this section, we introduce the computational methods (algorithms) used to select the optimal (best of breed) accessibility indices, whose correlation with the residuals of a naïve ordinary least squares housing price model excluding locational attributes is the highest. Based on the assumption that these indices will be able to capture the effect of the locational attributes, we then further improve their definition by obtaining their orthogonal (boosted) representation through principal components analysis, whose meaning can be clearly interpreted through their corresponding loading factors.

Hence, the new methodology proposed to calculate the orthogonal (or boosted) accessibility indices A_{nk}^{tm*} starts out by selecting the best accessibility index for each one of $K = 26$ variables presented in Table 2. The selection is made from the set of raw indices, A_{nk}^{tm} , based on the $Opportunity_{i(k)}(X, Y, d_{i(m)})$ features, that are defined for alternative values of β_m . Subsequently, as described below, we obtain a set of $k = 1, \dots, K$ optimal accessibility indices A_{nk}^{tm*} by choosing the β_m that maximizes the correlation with the errors of a naïve OLS model that omits the accessibility variables but includes the rest of attributes ($S_{nj}^t, M_{nl}^t, D_{nk}^t$). Then, these optimal indices are transformed into a set of orthogonal accessibility indices using principal components analysis. This transformation reduces the need for covariate treatment by removing collinearity among the indices, thereby improving the performance of the final regression model.

The whole process takes place in three steps that are programmed in two algorithms described in Fig. 3. The first one creates the family of candidate raw accessibility indices whereas the second is in charge

of selecting the best performing ones by an heuristic approach, and subsequently, perform the principal components analysis.⁴

4.1. Algorithm 1 - generating the locational seeds and candidate raw accessibility indices (A_{nk}^{tm})

Once we determine the *seed locations* following the procedure described above, we define the associated isochrone areas. For each location seed we have a set of isochrone areas for the series of time-distances: $d_{i(m=car)}$ and $d_{i(m=walk)}$. Fig. 4 illustrates the concentric rings defined around a specific seed.

Then, for each one of the rings we calculate all $Opportunity_k(X, Y, d_{i(m)})$ indices and consolidate them into the location accessibility index using the following values of β_m : 0.005, 0.01, 0.025, 0.05, 0.25, 0.5, 0.75, 1, and 2. Hence we obtain a family of 9 raw accessibility indices as presented in Eq. (2), each for a β_m value. That is $A_{nk}^{tm} = \sum_{i(m)=1}^I Opportunity_{i(k)}(X, Y, d_{i(m)}) \cdot e^{-\beta_m \cdot d_{i(m)}}$, $m = \{\text{car, walk}\}$, $\beta_m = 1, \dots, 9$. To calculate these indices we implement the following algorithm:

4.2. Algorithm 2 - selection of optimal (A_{nk}^{tm*}) and orthogonal (\hat{A}_{nk}^{tm*}) accessibility indices for hedonic price models

With all these candidate indices, 9 per opportunity variable, we initiate our second algorithm to determine which one would be the best performing in the hedonic price model (1). As aforementioned, to avoid

⁴ The Python scripts for the algorithms are available from the authors upon request.

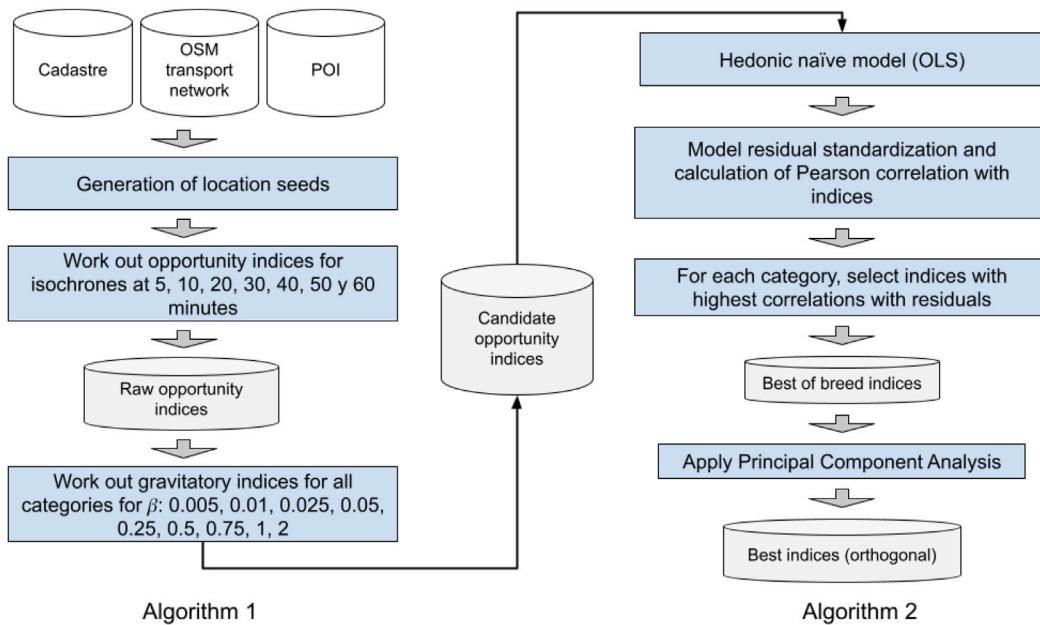


Fig. 3. Complete workflow for algorithms 1 and 2.

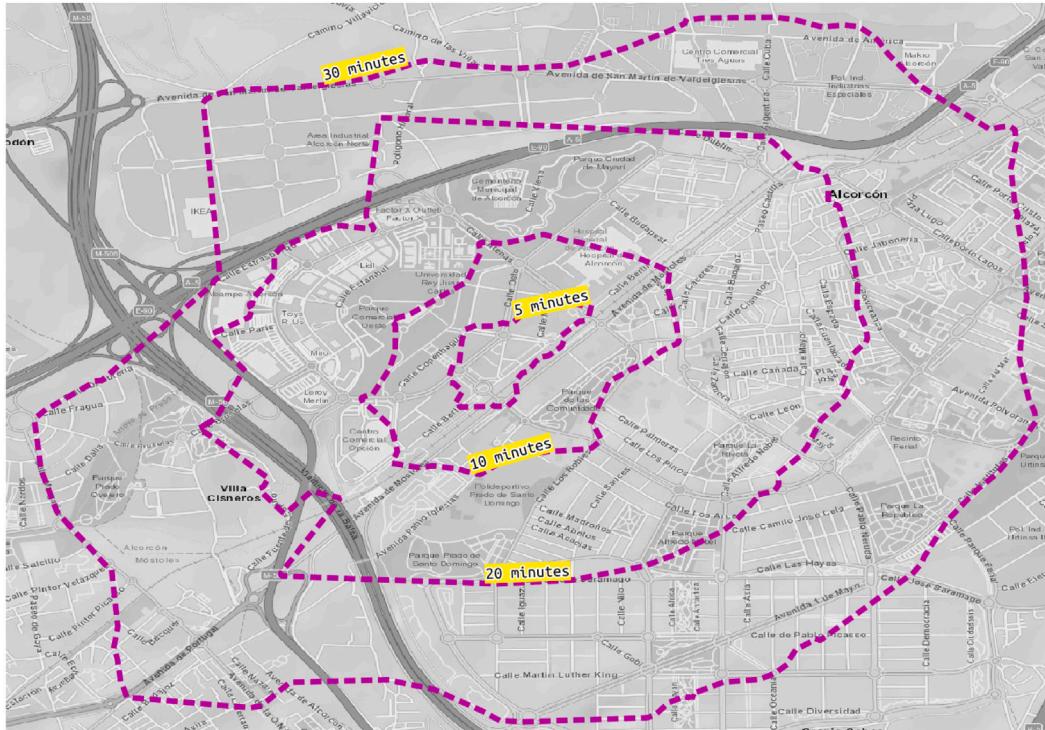


Fig. 4. Isochrone concentric rings are the areas reachable on foot at 5, 10, 20, 30 min from a location seed.

the evaluation of all combinations of variables and configurations, we follow a univariate heuristic approach consisting of selecting the β_m that yields the greatest correlation with the residuals of a naïve hedonic price model calculated without any locational information, but including the rest of the structural, market and time dummy variables: $(S_{nj}^t, M_{nl}^t, D_{nk}^t)$. The statistical rationale is the following: without any of the locational variables presented in Table 2, the residuals of such naïve model will show a high degree of spatial correlation (hypothesis corroborated in our empirical tests in section 6.2 below). Given this spatial non-stationarity, it is certain that any variable correlated with the residuals would be also correlated with the omitted spatial attributes,

hence being a good candidate as a predictor for the model. The heuristic approach makes a selection of the best performing configuration for each characteristic with the expectation of reducing the spatial bias of the naïve model. This procedure represents a simplified version of a boosting algorithm, see Friedman (2002).

Since the set of optimal accessibility indices A_{nk}^{tm*} may be correlated to each other, we finalize the second algorithm performing a Principal Component Analysis (PCA) to remove collinearity among variables (Abdi & Williams, 2010). Multicollinearity does not reduce predictive power yet it is detrimental to statistical inference by making the predictor coefficients and the R^2 measure less reliable (Orford, 2017).

Data: Cadastral Data, Open Street Maps network and POI
Result: Gravitational accessibility candidate Index
Generate seeds from coordinates of cadastral parcels with residential use;
forall Location Seeds do
| Generate isochrones for: 5, 10, 20, 30, 40, 50, 60 minutes;
forall Isochrones do
| **forall Opportunity Index Category do**
| | Generate Opportunity index value;
| **end**
| **end**
forall Opportunity Index Category do
| **forall** β_m in 0.005, 0.01, 0.025, 0.05, 0.25, 0.5, 0.75, 1, 2
| | **do**
| | | Create gravitational accessibility Index for Opportunity Index and β_m ;
| | **end**
| **end**
end

Algorithm 1: Part I - Generation of the family of raw accessibility indices: $\sum_{i(m)=1}^I Opportunity_{i(k)}(X, Y, d_{i(m)}) \cdot e^{-\beta_m \cdot d_{i(m)}}, m = \text{car, walk, } \beta_m = 1, \dots, 9.$

Table 5
Principal Components loadings - Walk mode.

Name	COMP	COMP	COMP	COMP
	WALK	WALK	WALK	WALK
	1	2	3	4
VACATIONAL	0.93	0.19		
TRANSPORT BUS	0.72	0.63		
TOURISM	0.92	0.23		
SHOP	0.85	0.44		
HOTEL	0.95	0.13		
FOOD	0.93	0.29		
CAD VENUES	0.91	0.35		
CAD RELIGION	0.77	0.59	0.12	
CAD PUBLIC	0.88	0.39	0.10	
CAD OFFICE	0.75	0.56	0.13	
CAD HOTEL	0.83	0.51		
CAD COMMERCE	0.70	0.68		
TRANSPORT METRO	0.54	0.74	0.15	
SPORT	0.18	0.86	0.10	
PARK	0.46	0.78		
HEALTH	0.64	0.70		
EDUCATION	0.52	0.80	0.16	
CAD SCHOOLS	0.57	0.77	0.15	
CAD RESIDENTIAL	0.58	0.76	0.13	
CAD INDUSTRY		0.85		
ROUTING COMPLEXITY	0.11	0.12	0.98	
CAD SPORT		0.17		0.97
TRANSPORT TRAIN	0.35	0.28		
ROUTING HIGHWAY	0.21	0.25		

Transformed variables retain the information from the original indices while conforming a set of new orthogonal variables called principal components. Consequently, the obtained accessibility components \hat{A}_{nk}^{tm*} are orthogonal among themselves, and therefore can be employed to estimate the hedonic housing price model without worrying about the likely multicollinearity of the raw accessibility indices.

In on-line Appendix C we present the best performing β_m^* for each location accessibility index. We observe that the car mode demands a stronger exponential decay with a median value equal to $\beta_{car}^* = 0.05$, whereas the median value for walking stands at $\beta_{walk}^* = 0.005$. These values imply a decay factor for a 10 min trip equivalent to 39.35% and 4.89%, respectively (calculated as $1 - e^{-\beta_m^* \cdot 10 \text{ minutes}}$). We illustrate our results in Fig. 5 showing the optimal accessibility indices

Data: Gravitational accessibility candidate Index

Result: Best performing gravitational indices (best of breed and orthogonal)

Create naïve regression OLS hedonic model ; Standardize model residuals ;

forall Gravitational accessibility candidate Index do

| Calculate Pearson Correlation Index between standardized residuals and candidate Index ;

end

forall Raw Opportunity Index do

| Select candidate index with the highest abs(correlation index) as Best-Of-Breed candidate;

end

Create Orthogonal Accessibility indices using Principal Components Analysis from best of breed indices

Algorithm 2: Part II - Selection of best of breed location accessibility indices

A_{nk}^{tm*} for selected variables. We show only results for walk indices given the European urban configuration of Madrid, characterized by a compact layout and high population densities—in comparison to American cities whose urban sprawl normally requires transportation by car. Lighter colors indicate greater access to opportunities. For example, hotels (WALK_HOTELS) are highly concentrated in the city center whereas Health services are more evenly distributed across the city (WALK_HEALTH).

4.3. Orthogonal (\hat{A}_{nk}^{tm}) accessibility indices

Finally, the principal components analysis yields 6 orthogonal accessibility indices (\hat{A}_{nk}^{tm}) for the walk and car transport modes, which respectively account for 94.7% and 98.3% of the variance of the respective raw accessibility measures—on-line Appendix D shows the principal components' scree tables for both transportation modes. The principal components are denoted by COMP_WALK_* and COMP_CAR_* in the regression results. Fig. 6 illustrates the three main accessibility components for walking (in terms of their eigenvalues).

The interpretation of these components follows the five categories presented in Table 2. To discern their meaning we calculate the PCA loadings, shown in Tables 5 and 6. The loadings measure the degree of contribution of each variable to the orthogonal accessibility indices. Also, to improve its readability we apply a Varimax rotation to the original values (Kaiser, 1958). This transformation, based on maximizing the variances of the squared loadings, maintains the original structure of data, although maximizing the distinction and differentiation of variable and factor correlations. Detailed results of the transformation method are provided in Tables A3 and A4 in on-line Appendix D, presenting the extraction and rotated sums of squared loadings. Again, the orthogonal (or boosted) accessibility indices obtained in this study after applying the optimization process and principal components analysis are available at <https://github.com/davidreyblanco/accessibility/tree/master>.

As we see in Table 5 presenting the raw indices' loadings, the first principal component COMP_WALK_1 refers to areas with a high degree of hotels, leisure services, commercial areas and well connected with public transport and with a high existence of touristic POIs. The second component highlights the immediate outer ring of the city center, affluent residential urban areas from a urban standpoint. Consequently, we see a lower correlation with touristic amenities, hotels and restaurants and offices but still well connected and with a high presence of commerce and residential areas. The third and fourth components do not contribute to reduce the dimensionality of the analysis as they are mainly related to a single variable. We see that they are related to access to areas with high road density (normally congested which

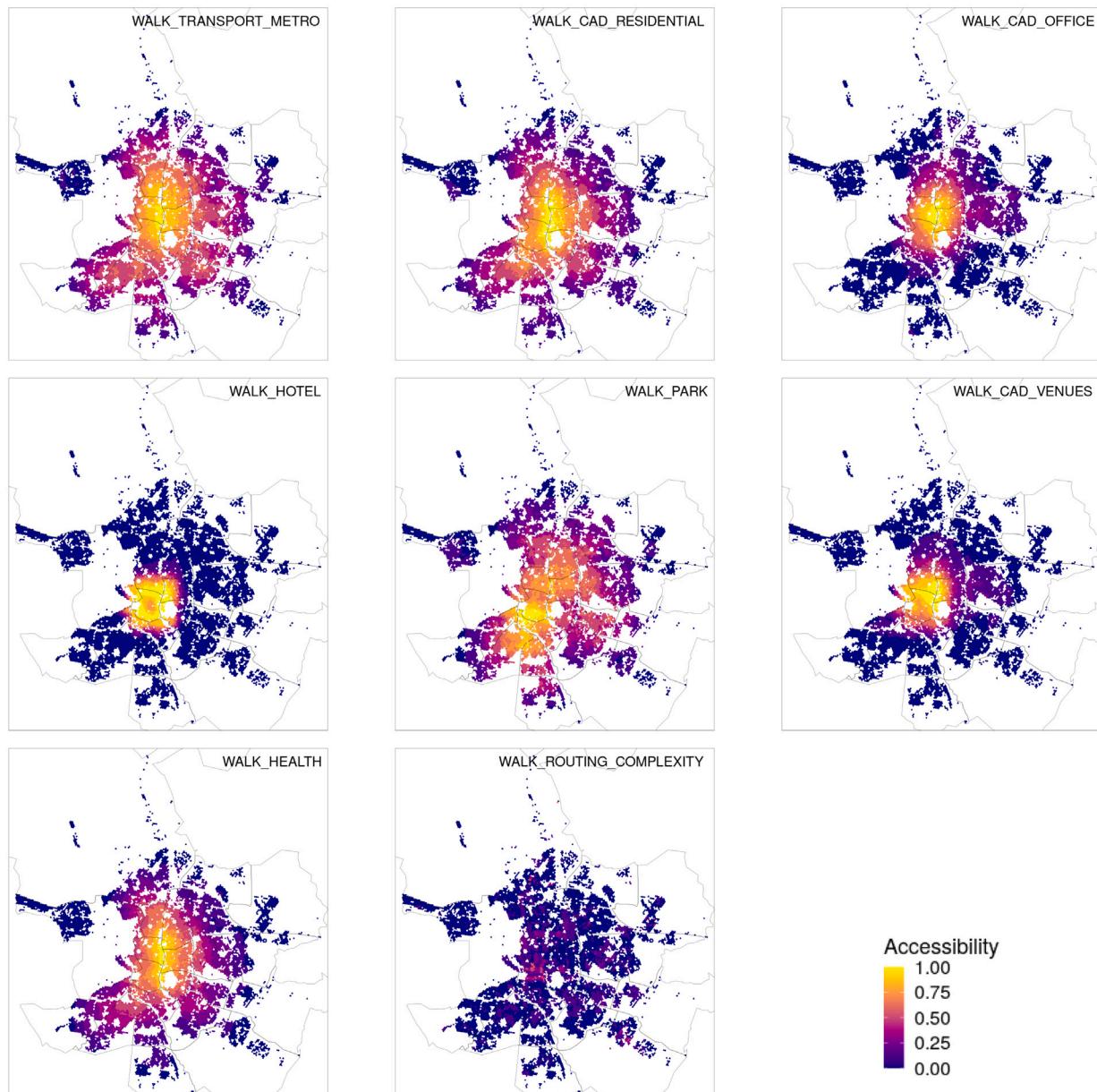


Fig. 5. Optimal accessibility indices (A_{nk}^{im*}).

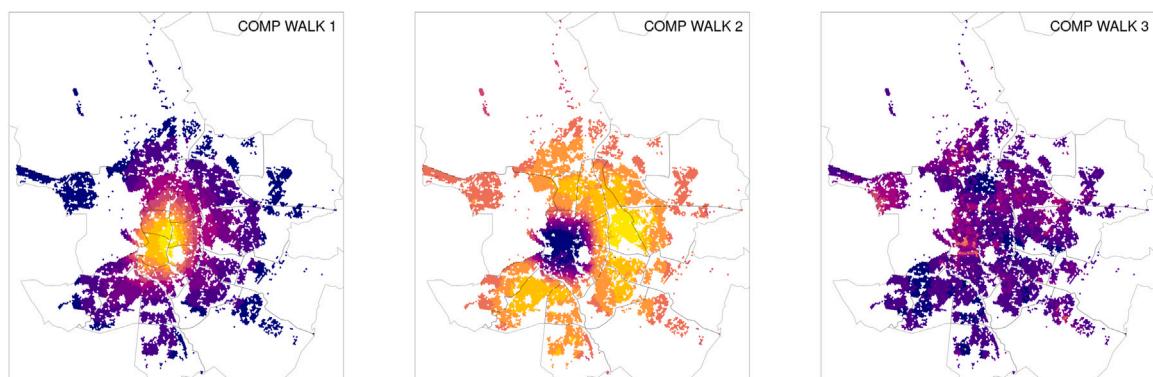


Fig. 6. Orthogonal accessibility indices \hat{A}_{nk}^{im*} for walk mode.

Table 6

Principal Components loadings - Car mode.

Name	COMP CAR 1	COMP CAR 2	COMP CAR 3	COMP CAR 4
TRANSPORT BUS	0.95	0.24	0.15	
TRANSPORT AIRPORT	0.71	0.56	0.15	
TOURISM	0.91	0.34	0.13	
SHOP	0.93	0.31	0.14	
HOTEL	0.96	0.19	0.11	
HEALTH	0.95	0.25	0.14	
CAD VENUES	0.94	0.30	0.15	
CAD URBAN LAND	0.81	0.45	0.17	
CAD SCHOOLS	0.90	0.37	0.15	
CAD RELIGION	0.93	0.32	0.14	
CAD PUBLIC	0.92	0.35	0.15	
CAD OFFICE	0.90	0.37	0.14	
CAD INDUSTRY	0.79	0.54	0.15	
CAD HOTEL	0.94	0.30	0.14	
CAD COMMERCE	0.92	0.33	0.15	
CAD AGRICULTURE	0.68	0.54	0.13	
TRANSPORT TRAIN	0.66	0.68	0.14	
SPORT	0.49	0.85	0.10	
PARK	0.50	0.85	0.11	
EDUCATION		0.97		
CAD SPORT	0.29	0.74		
CAD RESIDENTIAL	0.49	0.85		0.11
ROUTING HIGHWAY			1	
ROUTING COMPLEXITY	-0.41	-0.20		-0.89

favors walking) and sport facilities, respectively. Focusing on the car transportation mode, [Table 6](#), we observe that the first one is also related to touristic and commercial areas, public transportation by bus, and office campuses and industrial areas normally found in the periphery of the city that are mainly accessible by car. The second component is related to some social, recreational (park, sport) and educational opportunities, while the third is fully correlated with the extent of the highway system, again accessible by car only. Interestingly, the fourth and final component captures the disutility associated to traffic congestion in European cities like Madrid in areas with a high concentration of roads, avenues, streets, etc., which are permanently clogged, thereby discouraging the use of private car. A similar degree of interpretability is found by [Čeh et al. \(2018\)](#) in their hedonic model for Ljubljana that applies PCA to all explanatory variables, but including standard distance-based accessibility to transport infrastructure, public services and POI.

5. Results: Regression and machine learning approaches

In this section, we put to the test our newly created orthogonal (boosted) accessibility indices by including them as explanatory variables in the standard hedonic housing price model (1). We estimate this model using different regression methods and machine learning techniques, and conclude their superiority over specifications that either disregard accessibility attributes, or include them though simple locational dummies. The increased accuracy (uplift) of the estimations including the new indices is measured through the usual error metrics.

5.1. Sample

For the implementation of the hedonic price model (HPM) and to check the robustness of the results, we consider three incremental models, each with the following variations of the dataset:

- **None:** Uses the original Idealista data presented in [Table 4](#) with all accessibility-wise variables removed. This constitutes the aforementioned *naïve* model. We assume this model as a baseline scenario to benchmark the performance of the model including the accessibility indices.

- **Dummy:** Following common practice, this specification models location through the inclusion of a binary variable for each district (location dummy), intended to capture the difference in price from one geographical area to another.
- **Accessibility:** The most comprehensive specification. It includes the orthogonal (or boosted) accessibility indices obtained through principal component analysis.

5.2. Hedonic housing price model algorithms

To estimate housing prices we rely on a range of modeling techniques, including regression models and machine learning models. As previously anticipated, for the econometric approach we use the standard OLS model to test if the orthogonal location indices perform well in terms of expected signs and statistical significance. In this regard we avoid the use of more complex approaches like local regressions. Given that the main goal of our study is to maximize the accuracy in predicting house prices, in the machine learning approach we include complex tree models. These models are capable of overcoming some of the limitation of regression models, although they are more difficult to tune and prone to over-fitting, especially in the case of boosting models.

We select a total of four models: the first two based on regression analysis and the last two on machine learning. For each approach there is a basic specification, followed by an improved method.

- **Ordinary Least Squared Regression, OLS.** Standard model estimating the parameters of a linear function by minimizing the sum of squared residuals.
- **Lasso and Elastic-Net Regularized Generalized Linear Models, LERG.** This method also estimates a linear regression model using a cyclical coordinate descent, computed along a regularization path—as implemented in the ‘glmnet R’ package ([Friedman, Hastie, & Tibshirani, 2009](#)). The method performs both a L1 (lasso) and L2 (ridge) penalties regularization. It is specially suited to our study given its large dimensionality and feature sparsity.
- **Recursive Partitioning Trees, RP Trees.** A basic regression tree model CART implemented from the ‘rpart R’ package ([Therneau, Atkinson, Ripley, & Ripley, 2022](#)). Originally proposed by Breiman in 1984 ([Breiman, 2017](#)), the model is built using a two stage procedure that results in binary decision trees.
- **Random Forests, R Forests.** This method is also a tree-based machine learning model that estimates the regression magnitude as a consensus of a number of models ([Breiman, 2001](#)). It operates by constructing a multitude of decision trees at training time and outputting mean prediction (regression) of the individual trees. This approach solves the overfitting problems that tree models exhibit (i.e., models do not generalize properly when replicating the training set behavior). It is also commonly used in the industry for HPM for its capacity to combine accuracy, generalization power and ability to capture non-linear relations. We use the ‘ranger R’ package by [Wright and Ziegler \(2015\)](#).

The four methods are estimated selecting as dependent variable unit prices (i.e., euros per square meter, held in the *UNITPRICE* variable). As anticipated, we have taken this target variable as it is less sensitive to heteroscedasticity and selection biases compared with the overall price of properties. Cross validation is chosen for assessing the model’s performance and generalization ability for three reasons ([Hastie, Tibshirani, & Friedman, 2017](#)): first, it reduces the bias of using a single training and testing set; second, it provides a more reliable estimate of the model’s generalization ability; and third, it helps identify overfitting. [Arlot and Celisse \(2010\)](#) stated that the optimal partition of the dataset (*N*-fold) ranges between 5 to 10. Thus, we decide to use a 5-fold configuration, enabling a proper balance between statistical and computational performance. Then data is shuffled and split into five

Table 7
OLS Regression coefficients.

	Estimate	Std.error	t value	Pr(> t)	Signif.
(Intercept)	3035.07	23.16	131.06	< 2e-16	***
CONSTRUCTEDAREA	-0.81	0.05	-14.90	< 2e-16	***
FLATLOCATION	159.70	4.41	36.24	< 2e-16	***
ROOMNUMBER	-160.94	2.21	-72.76	< 2e-16	***
ISSTUDIO	-57.70	16.67	-3.46	5.38e-04	***
ISPENTHOUSE	443.48	6.90	64.25	< 2e-16	***
HASLIFT	508.33	4.24	119.92	< 2e-16	***
MAXBUILDINGFLOOR	20.32	0.63	32.18	< 2e-16	***
HASANNEX	249.37	2.68	93.17	< 2e-16	***
COMP_WALK_1	324.17	1.87	173.26	< 2e-16	***
COMP_WALK_2	58.61	1.00	58.75	< 2e-16	***
COMP_WALK_3	111.91	1.90	58.81	< 2e-16	***
COMP_WALK_4	168.62	1.77	95.35	< 2e-16	***
COMP_CAR_3	-736.53	19.97	-36.89	< 2e-16	***
RENTSALE_RATIO	139.23	6.04	23.07	< 2e-16	***
ONMARKET_SALE	4191.73	115.98	36.14	< 2e-16	***
ONMARKET_RENT	3139.32	81.44	38.55	< 2e-16	***
DEMAND	-27.69	0.25	-111.29	< 2e-16	***
PERIOD_2018_01_31	-353.88	7.35	-48.14	< 2e-16	***
PERIOD_2018_02_28	-306.01	7.39	-41.39	< 2e-16	***
PERIOD_2018_03_31	-258.98	7.30	-35.49	< 2e-16	***
PERIOD_2018_04_30	-210.42	7.36	-28.59	< 2e-16	***
PERIOD_2018_05_31	-163.17	7.37	-22.15	< 2e-16	***
PERIOD_2018_06_30	-129.71	7.31	-17.74	< 2e-16	***
PERIOD_2018_07_31	-86.55	7.24	-11.96	< 2e-16	***
PERIOD_2018_08_31	-49.55	7.21	-6.87	6.28e-12	***
PERIOD_2018_09_30	-35.90	7.38	-4.87	1.14e-06	***
PERIOD_2018_10_31	-10.67	7.23	-1.48	1.40e-01	
PERIOD_2018_11_30	0.04	6.85	0.01	9.96e-01	
CADASTRALQUALITYID_1	509.35	25.55	19.93	< 2e-16	***
CADASTRALQUALITYID_2	272.94	19.16	14.25	< 2e-16	***
CADASTRALQUALITYID_3	92.66	17.56	5.28	1.32e-07	***
CADASTRALQUALITYID_4	-126.82	17.23	-7.36	1.84e-13	***
CADASTRALQUALITYID_5	-358.96	17.33	-20.71	< 2e-16	***
CADASTRALQUALITYID_6	-363.26	17.40	-20.88	< 2e-16	***
CADASTRALQUALITYID_7	-371.33	17.78	-20.89	< 2e-16	***
CADASTRALQUALITYID_8	-433.18	21.10	-20.53	< 2e-16	***

equally sized folds to be used in the corresponding experiments: one split is reserved for testing and the rest for training. Subsequently, we build and test the model five times and average the results of the five experiments. Performance measures are calculated as the average of the five executions.⁵ Finally, each algorithm is tuned by a parametrization analysis called grid search. This process tests several configurations and selects the best performing parameter sets. More details on the parametrization of the algorithms are provided in on-line Appendix E.

Before performing the 5-fold cross-validation of the models, Table 7 reports the results of the standard OLS approach for the whole sample. This allows us to show the basic relationship between the explanatory variables and house prices. Given the urban configuration of Madrid, favoring walk accessibility and public transportation, we rely on the first four walk accessibility indices presented in Table 5. The accessibility variables captured by these components make it unnecessary to include their car counterparts that are related to the same variables: hotel, tourism, commercial, etc. (i.e., walk and car components are positively correlated capturing the same information). The only exception is the third car component that is highly correlated with accessibility to highways and is not adequately represented by the walk components. Indeed, this component preserves the orthogonality constraint among accessibility covariates, given it is the least correlated with the walk components (displaying an average Pearson correlation coefficient to walk components of 0.16 in absolute terms). Moreover,

using the first four components for walk mode plus the third for car mode we jointly account for 92% of the variance. Results show that the coefficient of each selected accessibility component exhibits the expected sign and is statistically significant. The first component, COMP_WALK_1, displays a positive correlation to price in the OLS regression coefficients. This is corroborated by the spatial layout of values of the component, as shown in Figure A1 of on-line Appendix F, presenting the spatial distribution of prices and Figure A2 depicting the spatial layout of the components (the lighter the color the higher the value). Therefore the city center has higher values as it comprises areas with a large concentration of touristic establishments, shops and public transportation stops. Similar analyses can be made for the remaining walk components. Interestingly, the car component presents a negative sign, reflecting the disutility that brings being closer to highways (that is, away from the city center). This simply captures that in Madrid's case, housing in the periphery of the city along the highway network is less expensive. The rest of the covariates related to structural, market and time dummy variables exhibit the expected signs and are all significant at the 0.1% level. It is worth noting that looking at the sign and decreasing values of the time dummies, a general increase in prices took place throughout 2018.

5.3. Comparing methods: performance metrics

Table 8 presents the different metrics that are used in the study to benchmark model performance. These metrics are common and provide a standard representation of accuracy and predictive power. The performance of each method is measured in absolute terms through the *Mean Absolute Error*, *MAE*, and the *Median Absolute Error*, *MedAE*.

⁵ To prevent biased results when sampling for cross-validation the dataset is grouped by unique advertisement identifier, as the same ad can be found in more than one monthly record.

Table 8

Measure	Formula
Mean absolute error	$MAE = \frac{1}{n} \sum_{i=1}^n e_i^2$
Median absolute error	$MedAE = median e_i $
Mean absolute percentage error	$MAPE = \frac{1}{N} \sum_{i=1}^N \frac{ e_i }{x_i}$
R Squared, R^2	$R^2 = 1 - \frac{\sigma_{resid}^2}{\sigma^2}$

It is also measured in relative terms by way of the *Mean Absolute Percentage Error*, *MAPE*. We also report the correlation coefficient, R^2 , and the models' *uplift* by calculating the reduction in the *MAPE* with respect to the *naïve* baseline model 'None'.

5.4. Which model performs best?

Table 9 summarizes the performance of each method with respect to the chosen metrics. Regardless of the estimation method, the model including the optimal accessibility indices (Accessibility) outperforms its locational dummies counterpart (Dummy). We see a reasonable good performance of the *naïve* specification with no specific location attributes (None), however it carries area information as it contains the variables related to the location, such as the structure of the property and market features for their districts (see the Idealista dataset in **Table 4**). It is noticeable the good performance of the regressions methods, OLS and LERG, meaning that these variables are properly capturing the influence of locational attributes on price. Note also, that more complex LERG even yields marginally better results. On their part, simple trees (i.e., RP Tree) are not apparently able of generalizing location interactions with this set of variables, most likely due to not being sufficiently complex enough to incorporate location variables in detriment of other significant variables in the algorithm, such as the structural or market ones. However, the use of random forests (R Forests) presents the best performance indicators by far. Notice that despite the reduced number of principal components used, and thanks to the optimizing algorithms applied to generate them, our models display similar or even higher accuracy levels than others in the literature considering a larger number of locational variables. For instance, the study for the city of Madrid by [Del Cacho \(2010\)](#), using real estate portal data but relying on a smaller sample (25,415 observations), obtains a mean percentage error of 15.25%.

6. Spatial robustness and spatial autocorrelation

In this section, we study the spatial robustness of the previous hedonic price estimations. Assuming that our newly proposed orthogonal (or boosted) indices capture the effect of locational attributes on prices, their estimated values should exhibit low spatial autocorrelation. We rely on cross-validation and calculate the usual error metrics and Moran's coefficient to conclude the superiority of the proposed methods.

6.1. How spatially robust is the use of orthogonal accessibility indices \hat{A}_{nk}^{tm*}

As the previous results could be spatially biased or present too much over-fitting we challenge the models with spatial cross validation. This method is similar to a regular cross-validation but considering also non geographically overlapped folds. We establish five areas, 5-folds, one held out for validation and the rest four being used to construct the model. The main implication of this approach is that models are constructed with data from different locations than those considered in the validation. Therefore if a model fits the validation data, then it can be concluded that its locational attributes are properly modeled through the accessibility indices.

Table 10 reports the results for the cross-validation exercise. Again, we measure the degree of improvement of each model by the uplifting precision obtained with respect to the *naïve* specification excluding locational attributes. Again, the best performance of the 'Accessibility' model is a sign of the capability of the orthogonal (or boosted) indices in capturing the effect of the different location interactions on house prices. Further confirming this result is the remarkable uplift obtained with the linear regression models (OLS and LERG), which is on par with that obtained using machine learning techniques (Random Forests). However, the fact that random forests does not display a greater advantage over linear models suggests the existence of spatial over-fitting. This implies that this model can adjust for the existing interaction between locational attributes, but it is incapable of modeling new patterns such as those found in the validation data. This inability to learn and provide a general framework to explain spatial interactions finally shows up in the form of higher mean absolute percentage errors. Therefore, a main takeaway of our comparison of methods is that for areas with new urban configurations it might be preferred to rely on more parsimonious (and simpler) linear models.

In addition, to check whether the model performs better or worse across the Madrid area, we also measure the goodness-of-fit at the spatial level using a *pseudo local R²* over the hexagonal H3 grid. This statistic compares the relationship of the model's residuals for each tile $e_{(X,Y)}$ divided by the global target variable's variance—where, again, (X, Y) refers to each tile centroid.

$$Pseudo\ local\ R^2 = 1 - \frac{\sigma^2(e_{(X,Y)})}{\sigma^2(price/m^2)} \quad (4)$$

As shown [Fig. 7](#), R^2 decreases when we include the orthogonal accessibility indices, being this reduction higher for the machine learning (Random Forests) model. We observe that the city center is most prone to yield lower R^2 . This result might be caused by the nature of this submarket, as it does not behave as a pure residential area when compared with the rest of markets in the city, resulting in higher variability in prices. The uses of residential properties in Madrid's city center are mixed, including not only residential, but also short-term vacation rentals and professional purposes. However, in general, we conclude that introducing accessibility indices results in a notable gain in the goodness-of-fit. Still most centric areas are subject to higher model bias in all cases.

6.2. Are orthogonal accessibility indices \hat{A}_{nk}^{tm*} effective in controlling for spatial autocorrelation?

To test the hypothesis that the use of orthogonal accessibility indices is capable of capturing the effect of location interactions on housing prices, we measure the degree of autocorrelation of models' residuals. If accessibility indices capture the influence of location they would turn our hedonic price models' residuals into a spatial stationary process. Expressed from another angle, if the models are perfectly specified in terms of the considered variables, they will capture the effect of geographical features on housing prices, and therefore the model residuals will be uncorrelated with locational attributes. In spatial patterns it is expected that nearby observations share similar characteristics while differing from those located farther away. We rely on Moran's I coefficient of spatial autocorrelation for the analysis ([Moran, 1950](#)) In its simplest form, Moran's index is calculated by assigning weights to neighboring observations: 1 for bordering locations, and 0 otherwise. These weights constitute the so-called neighboring function, which can be defined in terms of proximity matrices that use different criteria (e.g. pair-wise distances among locations). Moran's index for a given variable of interest x is defined as follows:

$$I = \frac{\frac{n}{S_0} \sum_{i=1}^n \sum_{j=1}^n w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (5)$$

Table 9
Regular cross validation with 5 folds benchmark.

Method	Model	Mae	Medae	Mape	R ²	Uplift
OLS	None	712.15	532.50	22.5%	0.68	
	Dummy	638.92	455.16	19.2%	0.72	10.3%
	Accessibility	619.73	446.61	18.8%	0.74	13%
LERG	None	711.94	531.94	22.5%	0.68	
	Dummy	638.76	454.96	19.2%	0.72	10.3%
	Accessibility	619.49	446.16	18.7%	0.74	13%
RP Tree	None	559.91	397.00	17.6%	0.77	
	Dummy	558.62	395.39	17.5%	0.77	0.2%
	Accessibility	534.86	381.36	16.9%	0.79	4.5%
R Forest	None	443.61	285.48	13.5%	0.85	
	Dummy	392.02	209.03	12%	0.85	11.6%
	Accessibility	347.68	176.56	10.6%	0.88	21.6%

Table 10
Spatial cross validation with 5 folds benchmark.

Method	Model	Mae	Medae	Mape	Uplift
OLS	None	1077.34	884.43	39.1%	
	Dummy	930.36	751.79	31.8%	13.6%
	Accessibility	704.67	561.28	24.1%	34.6%
LERG	None	1077.43	884.45	39.1%	
	Dummy	999.60	824.16	35.7%	7.2%
	Accessibility	703.92	560.15	24.0%	34.7%
RP Tree	None	1187.67	1010.30	43.2%	
	Dummy	1079.59	934.13	36.8%	9.1%
	Accessibility	842.44	679.91	28.3%	29.1%
R Forest	None	1078.52	892.49	38.6%	
	Dummy	998.55	822.26	33.8%	7.4%
	Accessibility	697.07	538.72	23.0%	35.4%

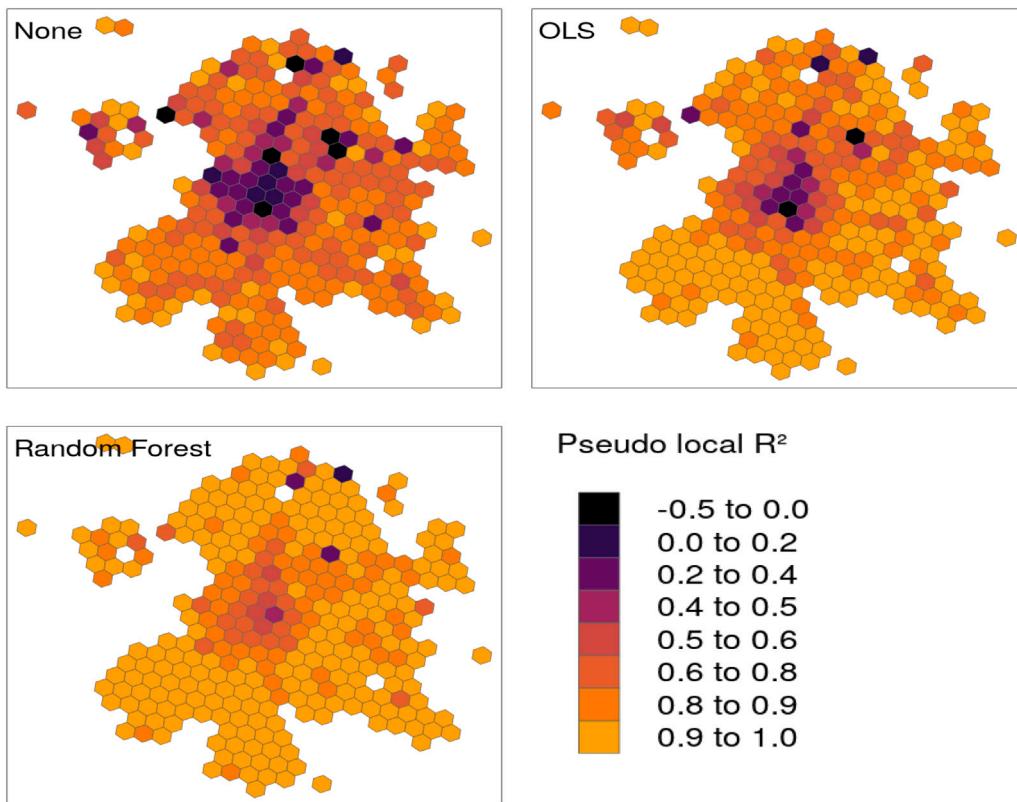


Fig. 7. Orthogonal accessibility for walk transport mode indices.

where w_{ij} is the weight between observation i and j , x_i and x_j are their respective variables of interest, and S_0 is the sum of all w_{ij} 's: $S_0 = \sum_{i=1}^n \sum_{j=1}^n w_{ij}, w_{ii} = 0$.

Our baseline for the comparison is the global autocorrelation of residuals estimated from the *naïve OLS* model excluding the location variables. The obtained global autocorrelation is 0.652, which reduces to 0.374 for the ‘Accessibility’ model including the orthogonal accessibility indices. For the Random Forests, the inclusion of the location indices reduces Moran’s I statistic to just about zero: 0.030. More detailed data is available in on-line Appendix G. We conclude that the optimal accessibility variables, calculated through our automated process, help reduce significantly the autocorrelation of the models’ residuals, overcoming the major drawback of having to define beforehand the accessibility covariates as in Morali and Yilmaz (2020) or Čeh et al. (2018).

7. Conclusions and future research

The results obtained confirm that our proposed methodology to incorporate qualified accessibility indices into hedonic housing price modeling, which are calculated through an automated process combining geographical information systems (used of discrete global grid system, DGGS) and statistical methods (i.e., principal components analysis), is capable of capturing the effect of locational attributes, thereby improving estimation performance, both in terms of goodness-of-fit and spatial correlation (Diewert & Shimizu, 2021). Our methodology aims at improving the predictions of real state prices for academic research, institutional use (e.g., statistical offices) and industry stakeholders. The use of optimal accessibility indices over a bandwidth of isochrones results in better performance than the traditional location dummy approach or the use of simple accessibility variables based on distance and time measures (Heyman & Sommervoll, 2019), while reducing notably the effects of spatial autocorrelation. The performance gain is seen with several modeling methods: it reaches a 13.0% accuracy uplift in the regressions, way higher than for recursive partition trees, 4.5%, but lower than for random forests, 21.6%. As for the performance gain in terms of spatial correlation, they are on par for the three models: 34.6% (OLS), 29.1% (RP Trees) and 35.4% (RF). Therefore we conclude that, compared to previous results in the literature where the superiority of machine learning methods for price prediction was undisputed, the use of our methodology can bring new arguments in favor of regression analyses, given their good performance and explanatory capacity. Nevertheless machine learning techniques like random forests are still capable of handling multiple types of location and price spatial interactions, especially nonlinear ones, see Rico-Juan and Taltavull (2021). Moreover, they overcome major limitations of hedonic regressions, which specify a single interaction rule for all areas, whereas the random forests model sets particular tree rules for the different areas, thereby adjusting these rules when needed.

Methodologically, the most important contributions of our study are: (1) the reduction of complexity in the definition of granular location attributes, which overcomes dimensionality problems, (2) the selection of the best gravity-based accessibility indices through an automated process, and (3) the use of principal components analysis to achieve orthogonality prior to their inclusion in the HPM. Regarding the dimensionality issue at the geographical level, the use of a discrete spatial mesh makes the algorithm highly efficient at the computational level and thus suitable for processes involving a great volume of price valuations in real time. Regarding the selection of the best accessibility indices, our algorithms choose the best gravitational specifications (in terms of decay patterns β_m) and, afterwards, the best-of-breed selection of the orthogonal (or boosted) accessibility indices. Key for the better performance of regression methods is that the optimal indices are uncorrelated with the residuals of an OLS model excluding spatial attributes. What makes this method relevant is that it can be applied to any type of real state (residential, commercial, industrial, raw land, and

special use) and scalable to any geographical dimension (municipal, regional, national, etc.).

Future research will be focused on several questions, first the use of other multivariate statistical methods besides principal components analysis to generate the optimal accessibility indices. Secondly, the possibility of evolving the univariate heuristic selection model to take into consideration the interaction among variables. Thirdly, bring other established machine learning techniques to the comparison (e.g., boosting, neural networks, …). Finally, to study the application of orthogonal accessibility indices to spatial regression models, as we have the insight that the use of this new set of variables, combined with spatial econometrics (e.g., geographically weighted regressions), may result in even better performance (Wonseok & Nam, 2019).

CRediT authorship contribution statement

David Rey-Blanco: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – original draft, Writing – review & editing. **José L. Zofío:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing, Formal analysis. **Julio González-Arias:** Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We have shared data through the following links: <https://data.mendeley.com/datasets/3zzz8m5p3m/1> and <https://github.com/davidreyblanco/accessibility/tree/master>.

Acknowledgments

The authors thank Pelayo Arbués and Alessandro Galesi from Idealista for their support in the elaboration of this paper. José L. Zofío acknowledges financial support from Ministerio de Ciencia e Innovación/Agencia Estatal de Investigación, Spain under research grant EIN2020-112260/AEI/10.13039/501100011033. This work is the result of a joint research project by Universidad Nacional de Educación a Distancia (UNED) and Idealista. It relies partly on data provided by Idealista/data (2019). Access to the data used in the study can be obtained by contacting the corresponding author or Idealista at <https://www.idealista.com/data>.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.eswa.2023.121059>.

References

- Abdi, H., & Williams, L. J. (2010). Principal component analysis. In *Wiley interdisciplinary reviews: Computational statistics*, vol. 2, no. 4 (pp. 433–459). Wiley Online Library.
- Agostini, C., & Palmucci, G. (2017). Capitalización anticipada del metro de santiago en el precio de las viviendas. *El Trimestre Económico*, 75, 403–431.
- Airbnb (2017). Inside airbnb - Get the data. <http://insideairbnb.com/get-the-data.html>.
- Alfaró-Navarro, J. L., Cano, E. L., Alfaro-Cortes, 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*, <http://dx.doi.org/10.1155/2020/5287263>.
- Anselin, L., & Griffith, D. A. (1988). Do spatial effects really matter in regression analysis? In *Papers in regional science*, vol. 65, no. 1 (pp. 11–34). Wiley Online Library.

- Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications*, 36(2), 2843–2852.
- Small, K. A., & Song, S. (1994). Population and employment densities: Structure and change. *Journal of Urban Economics*, 36(3), 292–313.
- Sohn, D. W., Moudon, A. V., & Lee, J. (2012). The economic value of walkable neighborhoods. *Urban Design International*, 17(2), 115–128.
- Steurer, M., Hill, R. J., & Pfeifer, N. (2021). Metrics for evaluating the performance of machine learning based automated valuation models. *Journal of Property Research*, 38(2), 99–129.
- Stewart, J. Q. (1947). Empirical mathematical rules concerning the distribution and equilibrium of population. *Geographical Review*, 37(3), 461–485.
- Tang, Y., Zhang, L., Min, F., & He, J. (2022). Multiscale deep feature learning for human activity recognition using wearable sensors. *IEEE Transactions on Industrial Electronics*, 70(2), 2106–2116.
- Therneau, T., Atkinson, B., Ripley, B., & Ripley, B. (2022). Package Rpart. Available Online: <https://cran.r-project.org/web/packages/rpart/index.html>. (Accessed 23 August 2023).
- Uber (2018). H3 A hexagonal hierarchical geospatial indexing system. <https://uber.github.io/h3/#/>.
- Valier, A. (2020). Who performs better? AVMs vs hedonic models. *Journal of Property Investment & Finance*.
- Vecchio, G., & Martens, K. (2021). Accessibility and the capabilities approach: A review of the literature and proposal for conceptual advancements. *Transport Reviews*, 41(6), 833–854.
- Waddell, P., Berry, B. J. L., & Hoch, I. (1993). Residential property values in a Multinodal Urban Area: New evidence on the implicit price of location. *The Journal of Real Estate Finance and Economics*, 7(2), 117–141.
- Wang, X., Li, X., & Wu, J. (2020). House price index based on online listing information: The case of China. *Journal of Housing Economics*, 50, Article 101715.
- Wee, G. P., & Vickerman, R. (2021). Transport modes and accessibility. In *International encyclopedia of transportation*, vol. 5 (p. 1). Elsevier.
- Wonseok, S., & Nam, H. K. (2019). Trade-off relationship between public transportation accessibility and household economy: Analysis of subway access values by housing size. *Cities*, 89, 247–258.
- Wright, M. N., & Ziegler, A. (2015). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. arXiv preprint arXiv:1508.04409.
- Xiao, Y. (2017). *Urban morphology and housing market*. Springer.
- Yates, S., & Miller, N. (2011). Residential land values and walkability. *Journal of Sustainable Real Estate*, 3, 23–43.
- Zhou, Z., Chen, H., Hong, L., & Zhang, A. (2021). The effect of a subway on house prices: Evidence from Shanghai. *Real Estate Economics*, 49(1), 199–234.