Methods to examine omitted variable bias in hedonic price studies

David Boto-García

Department of Economics, University of Oviedo

botodavid@uniovi.es

Abstract:

Many studies use the hedonic pricing method to uncover consumers' willingness to pay for accommodation characteristics in the hospitality industry. In most empirical applications using cross-sectional data, implicit prices might be biased if omitted time-invariant variables correlate with observed attributes. This paper proposes a set of diagnostic checks to inspect the potential bias in the estimates through exploiting repeated information for the same accommodations. Due to the limitations imposed by standard fixed and random effects panel regressions, we advocate for the use of Mundlak and Hausman-Taylor estimators. The proposed methods are applied to a study of Airbnb hedonic prices in Ibiza using a 14-month time window. In doing so, we document a novel finding: Superhosts set *lower* prices conditional on time-invariant quality.

Keywords: *hedonic pricing; omitted variable bias; diagnostic checks; Airbnb; Superhost badge*

Cite as:

Boto-García, D. "Methods to examine omitted variable bias in hedonic price studies". *Tourism Economics*, forthcoming. https://doi.org/10.1177/13548166221113437

1. INTRODUCTION

The hedonic pricing method developed by Rosen (1974) has become a commonly used tool for uncovering tourists' preferences for accommodation attributes. By regressing accommodation prices on its characteristics, the researcher can estimate the implicit market price of a particular attribute holding everything else constant. This information is highly valuable for hospitality managers, since it allows the assessment of tourists' willingness to pay (henceforth WTP) for a marginal improve in a given characteristic.

A large body of research has conducted hedonic studies to identify the implicit prices of hotels (e.g., Abrate and Viglia, 2016), hostels (e.g., de Oliveira Santos, 2016) or Airbnb listings (e.g., Wang and Nicolau, 2017). However, hedonic regressions suffer from several caveats that can produce important distortions in the recovered shadow prices. Recently, Faye (2021) discusses important methodological aspects to bear in mind like the appropriate functional form, preference heterogeneity or time/spatial effects. Similarly, Boto-García (2022) illustrates how ignoring cross-sectional dependence in the form of spatial clusters overstates standard errors. Another usual problem is that the researcher cannot observe all the accommodation characteristics that are relevant to guests. Indeed, many quality factors (typically time-invariant) like décor, views or homeware are likely to be unobserved from the econometrician perspective. When these omitted variables correlate with observed attributes, standard OLS estimates of implicit prices are biased.

Empirical evidence on the magnitude of the marginal WTP for accommodation characteristics is rather mixed. Studies using data for several cities or regions show relevant spatial heterogeneity (Benítez-Aurioles, 2018; Gibbs et al., 2018; Wang and Nicolau, 2017; Moreno-Izquierdo et al., 2019; Gyodi and Nawarro, 2021). That is, consumers' preferences depend on the destination being analysed. As such, when the researcher encounters a result that is contrary to existing evidence (either in sign or magnitude), there is the doubt of whether it reflects a real pattern in the case study analysed or is a flaw caused by omitted characteristics. Therefore, the bias from omitted factors might produce misleading implications.

This paper proposes a set of diagnostic checks to be implemented for examining the magnitude of omitted variable bias from time-invariant attributes in hedonic price studies in tourism research.[1] The procedures can be also applied to other settings. When repeated price information for the same accommodation is available, panel data methods can be implemented

to explore the bias of standard cross-sectional OLS. Fixed effects (hereafter FE) regressions control for any unobserved accommodation characteristic but at the cost of not allowing to recover the implicit price of any time-invariant attribute. Random effects (hereafter RE) regressions allow for time-invariant characteristics together with unobserved random effects, but they impose the strong assumption that unobserved attributes are uncorrelated with observed characteristics. This assumption is unlikely to hold in most applications. To circumvent these drawbacks, we first propose a battery of tests and diagnostic checks to examine whether there is a problem of omitted variable bias in the model results. If this is the case, we then advocate for the use of Mundlak correlated random effects (Mundlak, 1978) and the Hausman-Taylor estimator (Hausman and Taylor, 1981) to tackle it. We discuss in detail the rationale for their implementation and their advantages over RE and FE models.

Next, we illustrate the usefulness of the diagnostic tools and estimators proposed using a case study of Airbnb hedonic price determinants in Ibiza (Spain). Using a panel dataset of 1,990 listings observed during 14 months, we estimate several hedonic regressions to show how the pooled OLS estimator produces biased estimates of the implicit prices of hedonic characteristics. Contrary to hotels in which quality differences can be controlled for using well-known star rating indicators, quality information is usually more deficient in Airbnb accommodations (Guttentag, 2015). Therefore, the bias from omitted time-invariant quality is even more problematic in the Airbnb context than in other settings.

The article has three distinctive contributions. First, it proposes a set of diagnostic checks to inspect if there is a problem of omitted variable bias from time-invariant attributes (i.e., quality) in the hedonic price function.[2] These tests can be easily implemented by practitioners in each case study. Second, it advocates for the use of panel datasets when estimating hedonic equations. Apart from the capacity to control for unobserved time-invariant factors, the use of repeated price information for the same properties allows recovering consumers' average marginal willingness to pay over time, averaging out seasonal effects that shift the hedonic function. Therefore, unlike the time specific WTP estimates provided by cross-sectional studies, panel hedonics inform about the full price function (Bishop and Timmins, 2018). Third, it illustrates how misleading could be the results from pooled OLS regressions. In particular, we show that, contrary to common wisdom and previous empirical evidence, hosts holding the *Superhost* badge are found to charge lower prices, *ceteris paribus*. Although this finding should

be confirmed in further empirical studies, we explain the reasons behind this pattern. In the lights of our results, practical and theoretical implications are discussed.

2. LITERATURE REVIEW

There is a vast empirical literature in tourism research that uses hedonic modelling to uncover the shadow prices of accommodation characteristics. Table A1 in Supplementary Material presents a summary of some selected studies. In general terms, hedonic studies in tourism consider five blocks of price determinants: (i) intrinsic characteristics/services provided, (ii) quality/reputation factors, (iii) rental conditions/policies, (iv) site-specific/locational factors, and (v) external market aspects.

The type of services/amenities provided and the room type are among the main price determinants for hotels (Rigall-i-Torrent and Fluvià, 2011; Sánchez-Lozano et al., 2021; Soler et al., 2019). In this vein, consumers are found to value the tenure of a swimming pool (Chen and Rothschild, 2010), indoor sports facilities (Saló et al., 2014) or room services (Rigall-i-Torrent et al., 2011). Other aspects like hotel size (Latinopoulos, 2018; Abrate and Viglia, 2016), meal plans (Sánchez-Lozano et al., 2021), hotel style (Soler et al., 2019), or sea view (Espinet et al., 2003; Fleischer, 2012; Latinopoulos, 2018) have also been shown to be relevant price predictors. In the case of Airbnb, entire apartments are more expensive (Deboosere et al., 2019; Falk et al., 2019) and prices linearly increase with the number of bedrooms and bathrooms (Benítez-Aurioles, 2018; Voltes-Dorta and Sánchez-Medina, 2020; Moreno-Izquierdo et al., 2020; Boto-García, 2022), which capture accommodation capacity. Daily rates also vary considerably by property type (Falk et al., 2019; Boto-García, 2022; Casamatta et al., 2022). The availability of car parking (Cai et al., 2019), wireless Internet (Wang and Nicolau, 2017), cable TV (Chattopadhyay and Mitra, 2019) or elevator (Chica-Olmo et al., 2020) are also typically associated with higher rates. Additionally, Airbnb prices are positively correlated with the number of photos (Moreno-Izquierdo et al., 2019; 2020; Gibbs et al., 2018; Casamatta et al., 2022), host response rate (Sainagui et al., 2021; Moreno-Izquierdo et al., 2020) and vary depending on the type of bed (Boto-García, 2022).

Quality-signalling factors are other important price predictors (Abrate et al., 2011). For the case of hotels, star rating is positively correlated with prices (Becerra et al., 2013; Chen and Rothschild, 2010; Rigall-i-Torrent et al., 2011; Saló et al., 2014; Abrate and Viglia, 2016;

Masiero et al., 2015). Other reputational aspects are online ratings (Latinopoulos, 2018), quality awards (Soler et al., 2019) or brand affiliation (Chen and Rothschild, 2010). In this regard, prices have been shown to be positively correlated with positive online reviews (de Oliveira Santos, 2016; Lawani et al., 2019) but negatively related to the number of reviews in the case of Airbnb (Lawani et al., 2019; Chattopadhyay and Mitra, 2019; Deboosere et al., 2019; Chica-Olmo et al., 2020; Moreno-Izquierdo et al., 2020; Voltes-Dorta and Sánchez-Medina, 2020). The latter could reflect problems of reverse causality if the number of reviews is a lower bound of demand. Experimental evidence presented in Ert et al. (2016) shows that host reputation according to online review scores and trustworthiness as perceived from photos positively affect the likelihood of selecting a listing, which translates into greater rates.

An often-neglected factor for explaining accommodation prices is seasonality. Most studies use data for a single period (Gibbs et al., 2018; Cai et al., 2019; Lawani et al., 2019; Chica-Olmo et al., 2020) or conduct separate analysis for off-peak and peak periods (Fleischer, 2002; Abrate et al., 2011; Voltes-Dorta and Sánchez-Medina, 2020). Therefore, their findings are very seasonal dependent. That is why a growing body of research is starting to consider longitudinal datasets to study price variations throughout the year. This stream of research documents that Airbnb listings are highly priced in the summer but become cheaper during the winter (Deboosere et al., 2019; Sainagui et al., 2021; Casamatta et al., 2022).

In the Airbnb market, scholars are starting to consider host characteristics in the hedonic price function. Listings that belong to hosts who hold the *Superhost* badge are generally found to be more expensive (Benítez-Aurioles, 2018; Gibbs et al., 2018; Cai et al., 2019; Moreno-Izquierdo et al., 2020; Gyodi and Nawarro, 2021; Voltes-Dorta and Inchausti-Sintes, 2021).[3]
However, others do not find significant differences (Sainagui et al., 2021) or even a negative effect on prices (Casamatta et al., 2022). Furthermore, the number of listings on property (as an indicator of professionalism) is a variable that is receiving increasing attention. Nonetheless, its relationship with prices is unconclusive. Whereas several studies find a positive association (Wang and Nicolau, 2017; Kwok and Xie, 2017; Moreno-Izquierdo et al., 2019; Chica-Olmo et al., 2020; Voltes-Dorta and Sánchez-Medina, 2020; Boto-García, 2022; Casamatta et al., 2022), others document professional hosts charge lower prices (Cai et al., 2019; Deboosere et al., 2019; Boto-García et al., 2022). Part of this unconclusive evidence seems to emerge from professionals offering better quality listings, as nicely illustrated by Arvanitidis et al. (2020).

Rental policies are another important price predictor. In the hotel industry, free cancellation and refund policies are generally associated with higher rates (Abrate and Viglia, 2016; Latinopoulos, 2018; Sánchez-Lozano et al., 2021). However, the opposite effect is found for the case of Airbnb. Wang and Nicolau (2017), Cai et al. (2019), Moreno-Izquierdo et al. (2019, 2020) and Boto-García (2022) find that a strict cancellation policy is associated with higher prices. Similarly, Benítez-Aurioles (2017, 2018) report that accommodations with flexible cancellation policies are cheaper. Concerning minimum stay requirements, empirical evidence is unconclusive (Moreno-Izquierdo et al., 2020; Sainagui et al., 2021). Other studies show that hosts who enable the 'instant book' feature in their listings (the listing is immediately booked without the need for approval from the host) are cheaper (Gibbs et al., 2018; Cai et al., 2019; Chica-Olmo et al., 2020).

The accommodation location is another major aspect for understanding prices. Indeed, many studies indicate their inclusion makes the greater improvement in model fit (Chica-Olmo et al., 2020). Tourists have strong preferences for being close to the destination sightseeing spots (Yang et al., 2018) so where the accommodation is located matters a great deal. Distance to the city centre has been the most used locational variable (Önder et al., 2019; Soler et al., 2019; Sainagui et al., 2021), although some studies consider the distance to transportation hubs (Deboosere et al., 2019; Boto-García et al., 2021; Gyodi and Nawarro, 2021), the closest beach (Latinopoulos, 2018) or major attractions (Önder et al., 2019; Cai et al., 2019). In general terms, the closer to points of interest, the higher the price. Given the usual spatial dispersion of Airbnb listings, other scholars have gone beyond and include socioeconomic characteristics of the neighbourhoods like median income, noise or the ethnic composition as price determinants (Rigall-i-Torrent and Fluvià, 2011; Saló et al., 2014; Moreno-Izquierdo et al., 2019; 2020; Cai et al., 2019; Deboosere et al., 2019; Chica-Olmo et al., 2020).

Another stream of research has started to consider spatial price dependencies by which prices are affected by the prevailing price levels in the surrounding area. Omitted environmental attributes make the error term to be spatially correlated across geographic areas (cross sectional dependence), rendering standard OLS estimates inadequate. For this reason, spatial econometric models that consider spatial lags, either of the dependent variable (Lawani et al., 2019; Boto-García et al., 2021), in the error term (Tang et al., 2019; Chica-Olmo et al., 2020), or more complex specifications (Gyodi and Nawarro, 2021), are becoming more used in hedonic price studies. Similarly, other scholars have implemented Geographical Weighted

Regression analysis by which the explanatory variables are allowed to have different effects depending on their geographic location (Zhang et al., 2011; Latinopoulos, 2018; Voltes-Dorta and Sánchez-Medina, 2020). An alternative approach is to include a set of spatial fixed effects in the model (i.e., dummies for subregions within the geographic area of study), as in Rigall-i-Torrent and Fluvià (2011), Saló et al. (2014), Sainagui et al. (2021), Casamatta et al. (2022). As shown by Anselin and Arribas-Bel (2013), spatial fixed effects capture spatial dependence with a group-wise or block structure.

Finally, the number and proximity of competitors have been shown to be positively associated with hotel prices (Balaguer and Pernías, 2013; Becerra et al., 2013), although some studies document non-significant effects (Abrate and Viglia, 2016). Tang et al. (2019) show that Airbnb prices are positively correlated with the number of listings within the zip code area but unrelated to the number of hotels. Similarly, Deboosere et al. (2019) document that Airbnb daily rates significantly increase with the number of listings in the census tract. Moreno-Izquierdo et al. (2020) find that Airbnb prices are positively associated with the ratio of regulated apartments to hotels.

3. UNDERPINNINGS, DIAGNOSTIC CHECKS AND HOW TO PROCEED

3.1.The hedonic price method

In this subsection, we characterize the traditional hedonic framework developed by Rosen (1974). This is a model of demand in a differentiated products market in which consumers maximize utility and firms/hospitality managers maximize revenues. Tourist accommodations are regarded as bundles of attributes so that consumers choose among combinations of characteristics à la Lancaster (1966). In our application, we consider the case of Airbnb listings. A detailed characterization for hotels is provided by Rigall-i-Torrent and Fluvià (2007).

Let a set of Airbnb listings indexed by *j*, for $j = 1, \dots, J$, to be completely described by a finite vector of attributes. Let X_j denote a set of time-invariant intrinsic characteristics like the number of bedrooms or the type of property that are observed by the analyst. Let ξ_j reflect time-invariant intrinsic attributes that are unobserved from the econometrician perspective but observed and valued by consumers in the market.[4] In addition, let us denote by ϑ_k a set of location fixed effects defined at some given geographical aggregation level *k* that capture time-

invariant neighbourhood factors that are valued by consumers like accessibility to transportation hubs, noise, security or air quality.[5]

In equilibrium, Airbnb prices in a point in time *t* can be expressed as a function of hedonic attributes so that $p_j = f(X_j,\ \xi_j, \vartheta_k)$, with $p(.)$ referring to the hedonic price function that maps the product characteristics and the Airbnb price. This price function is generally assumed to be linear in parameters so that the hedonic equation is given by:

$$p_j = \alpha + \beta X_j + \gamma \vartheta_k + \delta \xi_j + \varepsilon_j \quad (1)$$

where ε_j is an error term with zero mean and constant variance that is uncorrelated with the regressors and captures idiosyncratic deviates from the deterministic price prediction. Since ξ_j is unobserved, in a cross-sectional dataset equation (1) becomes:

$$p_j = \tilde{\alpha} + \tilde{\beta} X_j + \tilde{\gamma} \vartheta_k + \omega_j \quad (2)$$

with $\omega_j = \delta \xi_j + \varepsilon_j$ being a composed error term.

3.2.Omitted variable bias

A key assumption in empirical applications is that the omitted attributes are mean independent of the observed characteristics (i.e. $E(\xi_j | X_j, \vartheta_k) = 0$). However, this is unlikely to hold; in most cases, higher values of desirable omitted attributes (ξ_j) tend to be positively correlated with higher values of desirable observed characteristics (X_j, ϑ_k). As such, the implicit prices of observed characteristics ($\frac{\partial p^*}{\partial X}$ and $\frac{\partial p^*}{\partial \vartheta}$) are upward biased. In particular, the omitted variable bias for a given attribute $x_j \subset X_j$ is given by:

$$\underbrace{\frac{Cov(p_j, x_j)}{Var(x_j)}}_{\tilde{\beta}} = \beta + \delta \tau_{x\xi} \quad (3)$$

where $\tau_{x\xi}$ is the vector of coefficients from auxiliary regressions of the elements of ξ_j on x_j (Angrist and Pischke, 2008). The greater the correlation between ξ_j on x_j, the greater the magnitude of the bias. If the researcher is interested in knowing the real consumer's willingness

to pay for an increase in attribute x_j (i.e. $\frac{\partial p^*}{\partial x_j}$), the estimate of $\tilde{\beta}$ in (2) could lead to misleading conclusions.

3.3.Diagnostic checks for empirical analyses

The omitted variable bias described before is a classical problem in hedonic price studies. Some of the proposals to tackle it require exploiting quasi-experimental designs with spatial/ temporal discontinuities in which changes in the variables of interest can be considered *as if* it were randomly assigned. Some examples are Greenstone and Gallagher (2008), Boes and Nüesch (2011) and Linden and Rockoff (2008). However, these identifying strategies are rare in common applications in tourism. Moreover, difference-in-differences settings make the dependent variable to be a change in prices (which mixes information about two distinct equilibria), and cannot inform about the implicit prices of time-invariant characteristics The only way to tackle omitted variable bias is to exploit information about prices and characteristics repeatedly over time (i.e., panel datasets). If the researcher has longitudinal data for the same properties over several periods *T*, for $t = 1, \dots, T$, panel data models can be estimated. In such cases, the time-invariant unobserved attributes ξ_j are explicitly modelled either as parameters to be estimated ('fixed' effects) or random variables drawn from a normal distribution ('random' effects) as follows:

$$p_{jt} = \alpha + \beta_1 X_j + \beta_2 Z_{jt} + \gamma \vartheta_k + \pi T_t + \xi_j + \varepsilon_{jt} \quad (4)$$

where T_t reflects time effects (yearly or monthly dummies) that capture shifts in price levels caused by seasonality or macroeconomic factors (inflation, changes in disposable income, etc.). The set of observed characteristics is here split into factors that are time-invariant (X_j) and factors that can change over time (Z_{jt}).

As mentioned before, panel hedonics capture consumers' mean WTP for characteristics rather than equilibrium conditions at a given point in time (Bishop and Timmins, 2018). That is, the partial derivatives exploit information about consumers' preferences along the price function by averaging seasonal effects.

If the unobserved attributes are independent from the observed characteristics, ξ_j can be treated as 'random' and equation (4) can be easily estimated using a random effects panel linear

regression by Generalized Least Squares (GLS). However, if $E(\xi_j|X_j, Z_{jt}, \vartheta_k) \neq 0$, the GLS RE estimator is inconsistent (Moulton, 1987). In such case, one can move to a panel fixed effects regression in which ξ_j are treated as parameters to be estimated, either using the within transformation or the least squared dummy variable estimator (LSDV). Nonetheless, this procedure has the important drawback that any time-invariant attribute is subsumed into the fixed effect so that the implicit prices of X_j cannot be recovered. In many applications, researchers are precisely interested in knowing the implicit prices of characteristics that are constant over time (e.g. type of property). As a result, researchers face a trade-off between avoiding omitted variable bias and estimating the implicit prices of time-invariant attributes. Even when working with panel datasets, most applications prioritize the latter (see Table A1) and estimate pooled OLS regressions with the corresponding risk of getting biased estimates. What can be done in this context?

The first step is to quantify the magnitude of the bias (if any) from omitted attributes. To this end, the following steps are proposed:

1. Run a pooled OLS regression of the price on the time-variant and time-invariant hedonic characteristics, the spatial fixed effects and the time dummies as in (4) ignoring the longitudinal nature of the dataset.
2. Run a (consistent) FE regression of the price on the time-variant attributes and the time effects treating ξ_j as parameters to be estimated. Then conduct a standard F test for Ho: $\xi_j = 0, \forall j$.
3. Run a RE regression of the price on the time-variant and time-invariant characteristics, the spatial fixed effects and the time dummies, treating ξ_j as random variables drawn from a normal distribution so that $\xi_j \sim N(0, \sigma_\xi)$.
4. Get the residuals from the pooled OLS regression in point 1. Then calculate the time means of the residuals for each unit j (i.e. $\overline{u_j} = \frac{1}{T_j}\sum_{t=1}^{T_j} \widehat{u_{jt}}$). This measure gathers the average deviation of the unit's residuals from the regression line expressed in units of the dependent variable and informs about the absolute size of the time-invariant unit unobserved effects (Moulton, 1987). A histogram of $\overline{u_j}$ can tentatively inform about the magnitude of the neglected unit-specific factors. A scatterplot of $\overline{u_j}$ against the fixed

effects estimates from point 2 (which share ξ_j in common) will also indicate whether there is something relevant hidden in the Pooled OLS regression.

5. Conduct a Breusch Pagan LM test (Breusch and Pagan, 1980) for random effects. The null hypothesis is that Var(ξ_j)=0. If we do not reject the null, then the pooled OLS will be consistent, and all the relevant hedonic attributes would be already considered. If it is rejected, then there is evidence that part of the residual variance comes from a listing-specific unobserved effect so that there are indeed time-invariant factors affecting prices that are not being modelled.
6. Conduct a Hausman test (Hausman, 1978) for choosing between the FE in point 2 and the RE model in point 3. The null hypothesis here is that FE (or alternatively, the LSDV estimator) and RE provide the same results (i.e., unobservables are uncorrelated with observed characteristics). If we do not reject the null, then we could run a RE panel regression as in (4) and get consistent estimates. However, if the null hypothesis is rejected, that would mean that the FE estimator is preferred: there is evidence that the listing-specific unobserved effects are correlated with explanatory variables. In this situation, knowing that the RE estimator is inconsistent, we can inspect the magnitude of the bias.
7. Compare the estimates for the common time-variant variables Z_{jt} between the RE and the FE and the OLS and the FE regressions. To this end, we can plot the point estimates of the time-varying variables for the FE and RE estimators and their standard errors to visually see how they depart one from another. This would help us to identify which variables are more biased. Additionally, the absolute relative bias could be calculated for each variable as $\frac{|\beta_{RE}-\beta_{FE}|}{\beta_{RE}}$ and $\frac{|\beta_{OLS}-\beta_{FE}|}{\beta_{FE}}$. Note the numerator of the relative bias captures the absolute value of the omitted variable bias in (3).
8. Inspect the bias in the OLS estimates for the time-invariant variables X_j. This consists of examining how sensitive a result is against the possibility of unobserved confounding factors. Suppose there is omitted variable bias, but the question is how large this bias could be. This can be done using alternative but similar procedures: (i) Oster's proportional selection in unobservables (Oster, 2019), (ii) Frank's sensitivity analysis (Frank, 2000) and (iii) Cinelli and Hazlett method to detect omitted variables (Cinelli and Hazlett, 2019).[6]

Suppose now we have done all these checks and we conclude that (i) there are unobserved listing characteristics that affect prices, (ii) these attributes are correlated with some of the explanatory variables, and (iii) a pooled OLS or a random effects model would produce biased estimates compared with a consistent FE estimator. As mentioned before, the problem with the within FE estimator is that it removes ξ_j and any time-invariant regressor X_j by mean-differencing the data prior to estimation so we cannot calculate $\frac{\partial p^*}{\partial X}$. In this case, valuable alternatives are the Mundlak correlated random effects regression (Mundlak, 1978) and the Hausman-Taylor estimator (Hausman and Taylor, 1981).

3.4.Mundlak correlated random effects

Mundlak (1978) propose a smart way to model the correlation between some explanatory variables and the unobserved listing-specific effect ξ_j in a RE context. The key assumption is that $E\left(\xi_j \middle| X_j, Z_{jt}, \vartheta_k, T_t\right) = \theta + \gamma \overline{Z}_j$, where $\overline{Z}_j$ is the time averages of Z_{jt}. Under this assumption, equation (4) can be rewritten as:

$$p_{jt} = \alpha + \beta_1 X_j + \beta_2 Z_{jt} + \gamma \vartheta_k + \pi T_t + \underbrace{\theta + \gamma \overline{Z}_j + \mu_j}_{\xi_j} + \varepsilon_{jt} \tag{5}$$

Since $E\left(\mu_j \middle| X_j, Z_{jt}, \vartheta_k, T_t, \overline{Z}_j\right) = 0$, the composed error term $\epsilon_{jt} = \mu_j + \varepsilon_{jt}$ is orthogonal to the explanatory variables and the estimation of (5) using a RE estimator provides consistent estimates.[7]

A Wald test for the null hypothesis that $\gamma = 0$ is a test for whether there is correlation between the unobserved effect and at least one time-varying variable. If this hypothesis is rejected, it means that at least one time-varying variable is correlated with the unobserved effect and the Mundlak approach is required for correct inference. If we do not reject the null, the model in (5) collapses to the standard RE regression, which in principle would produce consistent estimates of the time-varying variables.

As opposed to the FE, the Mundlak approach allows us to estimate β_1. However, we cannot rule out omitted variable bias in the time-invariant characteristics because this procedure only controls for correlation between ξ_j and the time-varying attributes Z_{jt}. Therefore, if we suspect

there is omitted variable bias in the time-invariant characteristics X_j, we need to move to the Hausman-Taylor (henceforth HT) estimator.

3.5.Hausman-Taylor estimator

Let $Z_{jt} = (z1_{jt}, z2_{jt})$ and $X_j = (x1_j, x2_j)$. The Hausman-Taylor estimator assumes a subset of the time-varying ($z1_{jt}$) and a subset of the time-invariant attributes ($x1_j$) are potentially endogenous (i.e., they are correlated with the unobserved attributes ξ_j). By contrast, $z2_{jt}$ and $x2_j$ are assumed to be exogenous. The procedure consists of first regressing prices on the time-varying attributes $z1_{jt}$ and $z2_{jt}$ using a within FE estimator. Second, the within residuals from this regression are then regressed on $x1_j$ and $x2_j$ using $x2_j$ and $z2_{jt}$ as instrumental variables, respectively. The overall and within residuals from the latter regression are subsequently used to estimate the components of variance of the dependent variable, which are then used to perform standard RE regression by GLS.

To be identified, the model needs more time-varying exogenous variables ($z2_{jt}$) than time-invariant endogenous variables ($x1_j$). Moreover, the instruments need to be sufficiently correlated with $x1_j$ to avoid a weak-instrument problem. If these two conditions are fulfilled, the HT estimator provides consistent estimates for both the time-varying and the time-invariant attributes.

4. DATA AND MODEL SPECIFICATION

4.1.Dataset and summary statistics

We use a rich panel dataset of Airbnb listings in Ibiza (Balearic Islands, Spain) obtained from AirDNA.[8] Ibiza is a well-known sun and beach destination, representing around 20% of the tourism flows to the Balearic Islands and which annually receives around 3 million tourists (IBESTAT, 2021). We have daily information for the prices and the property status (available, reserved or blocked) for 14 months, starting on 1st August 2015 and ending on 30th September 2016. Information about a set of listing intrinsic characteristics together with host-related aspects and rental rules is also available.

We work with monthly prices calculated as the average daily rate (in euros). This price does not include cleaning fees or additional charges for guests that are not included in the overall price. Figure 1 presents a histogram of monthly prices. Similar to other applications, average daily rates are highly skewed.

FIGURE 1 HERE

In total, we have valid data for 1,990 different listings located in 5 different municipalities and 19 distinct postal code areas (11,939 observations). Consistent with the literature on Airbnb pricing reviewed in Section 2, we consider the following blocks of hedonic price determinants:

- *Intrinsic characteristics*:
    - Type of property: dummy variables for Apartment, House, and Villa. The reference category gathers other types of properties like Bungalow, Chalet, Condominium, Dorm, Loft or Townhouse.
    - A binary indicator for whether the renting is for the entire property (as opposed to shared or private rooms)
    - Number of bedrooms
    - Number of bathrooms
    - Number of photos
    - A binary indicator for whether the listing is ready to host business travellers.[9]
- *Host characteristics*:
    - The number of listings the host owns in the island, as a proxy of professionalism.
    - A binary indicator for *Superhost* badge by Airbnb platform.
- *Rental rules*:
    - Minimum stay (in days)
    - A binary indicator for whether instant booking is enabled.
    - Cancellation policy: we consider two dummies for flexible and moderate cancellation policy, leaving strict as the reference category.
- *Postal code fixed effects*: as discussed in Section 2, many studies include environmental variables like the sociodemographic composition of the neighbourhood, accessibility to points of interest and transportation hubs, or ambient factors. To capture these aspects, we include a set of postal code fixed effects (dummy variables).
- *Competitive rivalry*:

- Number of Airbnb competitors: since Airbnb is considered as a competitive monopoly market (Boto-García et al., 2021; Casamatta et al., 2022), prices might be affected by the degree of competence in the area. Similar to Tang et al. (2019), we compute the mean number of Airbnb listings other that the one analysed in the neighbourhood per month. This is based on evidence presented in Voltes-Dorta and Inchausti-Sintes (2021) showing that price competition takes place at narrow spatial boundaries. Since the attractiveness of the area (locational quality) is controlled by the postal code fixed effects, the greater the supply of listings in the area, the lower the expected price, *ceteris paribus*.
- Number of hotel beds: also as in Tang et al. (2019), we include the number of hotel bed places per municipality and month. This information is retrieved from the Statistical Office of the Balearic Islands (IBESTAT). A negative relationship with prices is also expected here based on previous literature (e.g., Önder et al., 2019).
- Number of reservations in the previous month: To control for listings' attractiveness from aesthetics or host trustworthiness, we consider the number of times the listing was reserved in the previous month (lagged to avoid simultaneity). This variable is close to the occupancy rate of the property used by Moreno-Izquierdo et al. (2019).

- *Temporal effects*: to capture seasonal effects, we include monthly fixed effects. The first month (August 2015) acts as the excluded category.

Table 1 presents summary statistics of the variables. The number of listings the host owns, the number of Airbnb competitors/hotel beds in the neighbourhood, and the number of reservations in the past month are the only variables that vary over time. The rest are time-invariant.

TABLE 1 HERE

4.2.Functional form

One unresolved issue in hedonic price studies is the appropriate functional form (linear, semi-log, log-log) that describes the relationship between prices and their hedonic determinants. Earlier discussions on this can be found in Cropper et al. (1988) and Halvorsen and Pollakowski (1981). Although the semi-log is the most used (see Table A1 in Supplementary Material), Faye

(2021) recommends performing Box Cox transformation tests as a check. Using Monte Carlo simulations, Cropper et al. (1988) show that in the presence of potential omitted variables the linear Box Cox transformation and other parsimonious forms like the linear, the log-linear and the log-log are preferred (lower mean percentage error in estimating the WTP). A Box Cox regression of the price on the variables described above produces an estimate of lambda equal to 0.034, which is statistically different from zero (available upon request). As discussed in Cameron and Trivedi (2009, p.94), since the point estimate is closer to zero than to one, this provides greater support for a log-linear model. The log transformation of the price variable offers the advantage that it makes it to be normally distributed and it allows a straightforward interpretation of the variations of the dependent variable and the subsequent computation of hedonic price indexes given its affinity with Tornqvist price index (Hill, 2013). All in all, we estimate a log-linear hedonic price function.[10] The partial derivatives with respect to continuous variables are interpreted as semi-elasticities. For the case of a dummy variable D, the price premium is given by: $\frac{\partial \ln P}{\partial D} = (exp(\beta) - 1) * 100.$

Before moving on, it is important to note that some studies have used semiparametric and fully nonparametric hedonic price functions that avoid linearization strategies as an alternative to log-linear or Box-Cox regressions. Nonparametric regression estimators face the curse of dimensionality problem when the vector of product attributes involves many variables. A common approach is to allow the price function to be nonparametric only for a reduced subset of the regressors (Anglin and Gençay, 1996; Bontemps et al., 2008). To inspect the robustness of our findings to the linear-in-parameters assumption of the log-linear specification, we perform panel fixed effects semiparametric regressions for the continuous time-varying attributes as formulated in Baltagi and Li (2002) in the Supplementary Material, Figure A2. As shown there, the linearity in the continuous right-hand side variables seems to be appropriate in this context, but it seems worthwhile to explore in each case study. Cross-sectional semiparametric hedonic price regressions following Robinson (1988) offer similar results.

5. ANALYSIS

Table 2 presents the coefficient estimates for several log-linear hedonic price regressions: pooled OLS (Column 1), RE (Column 2), FE (Column 3), Mundlak (Column 4) and Hausman-Taylor (Column 5). Standard errors are clustered at the postal code level to acknowledge potential cross-correlation in residuals within geographic areas (Boto-García, 2022).

TABLE 2 HERE

Starting with the standard OLS estimates, the results are similar to previous studies.[11] We document that apartments are significantly cheaper than other property types (-25%), with no price differences between houses, villas and the other lodgings gathered in the reference category. Entire listings are highly priced, with rates being positively associated with the number of bedrooms (+0.17%) and bathrooms (+0.20%). Listings with a flexible cancellation policy are significantly more expensive (+9%). Similar to Benítez-Aurioles (2018), properties that allow for instant booking are cheaper (-7.7%). Holding the *Superhost* badge is associated with higher prices (+8%), although the coefficient is only significant at 90% confidence level. However, there is no association between prices and the number of photos, minimum stay and business ready certification. Strikingly, prices are unrelated with the number of listings the host has on property in the island. In line with Voltes-Dorta and Inchausti-Sintes (2021), prices are negatively related to the number of Airbnb listings in the neighbourhood (-1% per a 10-listing increase). However, as in Tang et al. (2019), there is no significant association between prices and the number of hotel beds in the neighbourhood. This finding is consistent with Voltes-Dorta and Inchausti-Sintes (2021) and suggests the negative effect of higher market supply could be offset by price complementarities through positive spillover effects. Furthermore, the number of days the property was booked in the previous month is negatively correlated with prices (-1.3% per day reserved). This suggests that listings that have been more demanded in the past charge lower prices. Finally, there is substantial price variability across months; Airbnb prices exhibit important seasonality, with hosts adjusting them depending on market conditions.

To first inspect the magnitude of the bias from omitted variables, we compute the time means of the pooled OLS residuals for each Airbnb listing ($\overline{u_j}$). Figure 2 reports a scatterplot of these mean residuals against the FE estimates of ξ_j from Column 3. The slope is 0.21. Note the FE incorporate both ξ_j and X_j whereas $\overline{u_j} = \xi_j + \overline{\varepsilon_j}$, with $\overline{\varepsilon_j} = 0$. We see there is a positive

association between the two: larger values of the listings' fixed effects positively correlate with larger values of the time means of residuals from OLS regression. Therefore, we can suspect there are relevant factors affecting prices that are neglected in the pooled OLS. Furthermore, the LM statistic for the presence of random effects takes value 24,015 (p-value<0.001) and the F test in the FE regression for the hypothesis that all $\xi_j = 0$ is 89.69 (p-value<0.001). The fraction of variance due to ξ_j in the composed error is also high (0.942). Altogether, there is evidence of time-invariant unobserved factors present in the data that need to be considered so that OLS estimates are likely to be biased.

FIGURE 2 HERE
FIGURE 3 HERE

Figure 3 plots the coefficient estimates for the time variant variables obtained from OLS, RE and FE. It seems there are not large differences between the FE and the RE. Since the differences appear not to be very large (see Table 2), one could think we could proceed with the RE estimates, which are more efficient and allow for time-invariant covariates. To formally check this, we conduct a Hausman test. However, the test rejects the null hypothesis that the differences between the FE and the RE estimates are not systematic (chi(17)=403.66, p-value<0.001). Furthermore, the correlation between the FE estimates (ξ_j) and the explanatory variables is -0.10. This suggests that the assumption that $E(\xi_j | X_j, Z_{jt}, \vartheta_k) = 0$ does not hold. Therefore, there is evidence that the RE estimates are inconsistent. However, as discussed before, the consistent FE estimates have the drawback they cannot inform about the implicit prices of time-invariant characteristics.

A F test for whether the time means of the time-variant variables are globally significant rejects the null hypothesis (chi2(17)=611.95, p-value<0.001), suggesting there is indeed correlation between the time varying regressors Z_{jt} and the random effects ξ_j. This reinforces the notion that the RE estimates are not consistent; Z_{jt} are not orthogonal to the composed error term. As such, this test favours the use of Mundlak estimates (Column 4) over RE (Column 3). Nonetheless, although the estimates for the time-varying variables Z_{jt} in the Mundlak regression are consistent, the implicit prices for the time-invariant attributes X_j might still be biased.

Next, we examine omitted variable bias in the pooled OLS regression. Table 3 below presents a sensitivity analysis based on the procedures developed by Cinelli and Hazlett (2019).[12] These authors propose two easy-to-understand measures that reparametrize the traditional omitted variable bias formula in (3) in terms of scale-free partial R^2 measures that capture the strength of the association between: (i) a potential confounder (e.g. unobserved quality, ξ_j) and each regressor $w \subset (W_{jt} = X_j, Z_{jt})$ conditional on the rest of variables, denoted by $R^2{}_{w\sim\xi|W}$, and (ii) the confounder and the dependent variable conditional on the rest of variables, denoted by $R^2{}_{Y\sim\xi|W}$. The corresponding formulas are provided in Cinelli and Hazlett (2019). Based on that, a robustness value (RV) can be constructed indicating how strong the association between a confounder and both the variable of interest and the dependent variable must be to reduce the estimated effect by 100% ($RV(\%)$) or to make the estimate not statistically significant at 95 confidence level ($RV_{\alpha=0.05}\ (\%)$). A RV close to 100 means that the variable of interest can handle strong confounders explaining almost all residual variation of the variable of interest and the dependent variable. By contrast, if the RV value is close to 0 is implies that very weak confounders could invalidate the OLS result.

A second measure proposed by Cinelli and Hazlett (2019) is the proportion of variation in the dependent variable that is uniquely explained by the variable of interest ($R^2{}_{Y\sim w|W}$), which gives a sensitivity analysis for an extreme scenario. If confounders explained 100% of the residual variance of the dependent variable, they would need to explain at least $R^2{}_{Y\sim w|W}(\%)$ of the residual variance of the variable of interest to bring down the estimated effect to zero.

TABLE 3 HERE

As presented in Table 3, these checks indicate that the point estimates and statistically significance of the pooled OLS regression are sensitive to uncontrolled quality, particularly the time-invariant observed covariates.[13] For instance, the RV value for *Apartment* is 11.52, which implies that unobserved confounders explaining at least 11.52% of the residual variance of both prices and *Apartment* would explain away the estimated 25.9% lower (exp(-0.293)-1) price of an apartment relative to the base category in the OLS regression in Table 2. Similarly, confounders would need to explain at least 9.92% of residual variance to make *Apartment* non-statistically significant at 95% confidence level based on $RV_{\alpha=0.05}\ (\%)$. Although variables like the number of bedrooms and bathrooms, the entire dummy or the number of reservations in past

month seem to be more robust to confounders, the implicit values for the cancellation policy, the business ready certification, the Superhost badge and the instant booking option seem to be more sensitive to uncontrolled confounders. That is why we finally move to the Hausman-Taylor estimator (hereafter HT).

We assume *Superhost* and the number of listings on property are two potential endogenous variables in the sense that their effect on prices is more likely to be confounded by unobserved quality. On the one hand, holding the *Superhost* badge implicitly gathers greater quality, since its concession is subject to the fulfilment of several requirements. The key issue is whether, compared to a property with a similar quality, Superhosts charge higher prices. For example, an unexperienced host might offer potential customers the same level of quality but without having the *Superhost* distinction yet (or offering other types of quality not considered among Superhost criteria). Indeed, the *Superhost* badge is not granted easily and only a small share of hosts attain it. On the other hand, studies on the differences in price setting between professional and non-professional hosts that use cross-sectional data have reported mixed findings (Wang and Nicolau, 2017; Kwok and Xie, 2017; Cai et al., 2019; Deboosere et al., 2019). We expect unobserved quality plays an important role because professionals' motivation to rent is profit maximization rather than peer sharing. As shown by some studies, the listings owned by professionals are better located (Xie and Mao, 2019) and are generally of better quality (Arvanitidis et al., 2020). As such, the non-significant effect documented in pooled OLS and RE regressions could be affected by this confounding.

A Sargan-Hansen test for whether the overidentifying restrictions for identification in HT does not reject the null hypothesis (chi(2)=20.91, p-value=0.139). Therefore, the estimates from HT are consistent and the most reliable in this context. This check is important for the validity of the estimator. We find that the estimates for the time-invariant attributes from the HT estimator differ notably from those from pooled OLS regression (column 1) and from the Mundlak model (column 4). Conditional on listing-specific effects, no price differences are now detected between apartments and the excluded category (gathering bungalows, chalets, condominiums, dorms, lofts or townhouses). The type of cancellation policy and the instant booking option turn to be non-significant. For the variables *Entire*, *Bedrooms* and *Bathrooms*, although a positive significant effect on prices is still documented, the effect size is slightly lower. Figure 4 compares the coefficient estimates for the time-invariant attributes (except Superhost, see below) across the different models. We see the coefficient estimates from OLS or RE

regressions are slightly biased. Concerning the time-variant attributes, the results are pretty similar to those from the Mundlak estimator.

FIGURE 4 HERE

Possibly the most shocking result is that *Superhosts* in the consistent HT regression are found to charge lower prices conditional on listings' quality. This finding is contrary to previous works, who agree to document a positive price premium for the badge (Benítez-Aurioles, 2018; Gibbs et al., 2018; Cai et al., 2019; Moreno-Izquierdo et al., 2020; Gyodi and Nawarro, 2021; Voltes-Dorta and Inchausti-Sintes, 2021). This puzzling evidence could be explained by omitted variable bias in the Superhost indicator, which partially captures listings' quality.[14] Existing literature use cross-sectional datasets in which Superhosts pricing strategies and their associated greater quality are confounded. Our panel analysis shows by contrast that, compared with a host that rents a property with similar quality, those who attain the *Superhost* badge set lower prices. Therefore, the estimates from pooled OLS or RE regressions for Superhost are likely to be severily biased, as illustrated in Figure 5.

FIGURE 5 HERE

Why do Superhosts charge lower prices, *ceteris paribus*? We argue this could be explained by these hosts acting professionally, reducing rates to ensure demand. To corroborate this, Figure 6 presents a scatterplot of the ADR on the number of reservations in the previous month. Red circles indicate listings with associated *Superhost* badge and blue diamonds refer to the rest. As can be seen, Superhosts set lower prices leaving aside quality aspects. Therefore, the common price premium for the *Superhost* badge reported in the literature could mainly reflect consumers' preferences over greater quality; conditional on that, Superhosts set lower daily rates.

FIGURE 6 HERE

6. CONCLUSIONS

6.1.Summary of findings

Hedonic price modelling using linear regression is a popular and theoretically consistent methodology to disentangle the implicit prices of the attributes embedded in differentiated products. Uncovering how much money (if any) people is willing to pay for a marginal gain in a certain individual characteristic has important implications for pricing strategies and managerial decision making. For this purpose, a large body of research has conducted hedonic price studies based on Rosen's framework in which the prices of tourism goods are regressed on their characteristics. Most applications in the hospitality industry have used cross-sectional datasets, typically due to data availability reasons. However, the results from cross-sectional linear regressions in this context are subject to omitted variable bias from uncontrolled time-invariant quality. When a study reports a finding that is contrary to previous evidence, there is the doubt of whether that is a real novel finding (given the case study context) or is the result of omitted confounding factors. One way to proceed is to use panel datasets in which the prices of the same goods are repeatedly observed over time and time-invariant heterogeneity can be controlled for.

We have discussed that standard fixed or random effects panel regressions do not solve the bias from omitted factors. This is because if the researcher is interested in assessing the hedonic value of a time-invariant attribute, its implicit value is not identified together with the fixed effects. Random effects linear regression is neither because of the restrictive assumption of uncorrelation between the individual effects and the time-invariant variables. That is why we advocate for the use of Mundlak correlated random effects (particularly when the variables of interest are time-variant) or the Hausman-Taylor estimator (mainly when the interest relies on time-invariant attributes). To illustrate the methods proposed, we have used a longitudinal data on Airbnb prices in Ibiza for 1,990 different properties during a 14-month window. We have estimated the hedonic equation using different econometric models and conducted a battery of different checks to examine the magnitude of omitted variable bias.

We have shown that the estimates for consumers' WTP are biased due to neglected time-invariant quality in cross-sectional OLS regressions. This seems to be particularly problematic for variables like the number of photos, business ready certification, the *Superhost* badge,

cancellation policies and the instant booking option. By contrast, the estimates for the structural characteristics like the number bedrooms and bathrooms, the type of building or the type of property, and time-variant regressors are less biased. Importantly, we have shown that, contrary to previous literature, Superhosts are found to charge lower prices conditional on quality. This could be the result of greater professionalism and profit maximization motivation among hosts that attain such badge as compared to hosts who offer listings of similar quality but without such distinction. This result highlights the importance of implementing appropriate econometric analysis to get accurate estimates of the implicit prices of accommodation characteristics.

6.2.Practical and theoretical contributions

This study makes methodological and theoretical contributions to the literature. Methodologically, we have implemented several econometric methods to deal with omitted variable bias. We have proposed a battery of tests that can be easily implemented in hedonic price studies by practitioners to inspect the sensitivity of their findings to omitted factors, not only in the Airbnb setting but also in other applications. If these checks suggest omitted variable bias is not present or it is of reduced magnitude, researchers can safely proceed with linear OLS regressions. However, if there is evidence of neglected heterogeneity in cross-sectional regressions, researchers must move to more sophisticated methods that try to avoid/ minimize the bias from omitted confounders. In any case, the reader must be aware that Mundlak and Hausman-Taylor estimator are not a panacea; their consistency strongly relies on the fulfilment of their identifying assumptions. Therefore, in situations when the use of panel datasets is not possible, we recommend researchers to report sensitivity analyses for the linear regression like the ones developed by Oster (2019), Frank (2000) or Cinelli and Hazlett (2019). When confounding cannot be rule out, it seems important to show how sensitive the documented results are to uncontrolled factors. Overall, we believe the methods explained in this article could be useful to draw more reliable estimates of consumers' WTP in hedonic price studies.

From a theoretical viewpoint, our analysis has shown that, once unobserved quality is controlled for, Superhosts set lower prices. This result is contrary to previous studies, who fail to distinguish between Superhosts pricing behaviour strategies and the fact their properties are of better quality. Because being granted such quality badge entails some degree of professionalism (as opposed to the casual peer-sharing motivation) in terms of keeping high occupancy, response rates, good ratings and low cancellation rates, it appears that Superhosts lower prices

to ensure demand. Although the price differential is estimated using data for a whole year averaging out seasonal effects, it is highly likely that the price differential becomes larger in the low season, when demand becomes more elastic and therefore price lowering becomes more effective for enhancing revenues. This finding expands our knowledge about pricing behavior in peer-to-peer markets but calls for more studies on the topic; we cannot rule out that this result is specific to the case study analysed. In any case, further research is needed to confirm that Superhosts set lower prices once quality has been controlled. Although there is abundant literature on accommodation pricing in general and Airbnb in particular, further research that deepens into compositional effects is still needed.

REFERENCES

Abrate, G., Capriello, A. and Fraquelli, G. 2011. “When quality signals talk: Evidence from the Turin hotel industry.” *Tourism Management* 32: 912-921.

Abrate, G. and Viglia, G. 2016. “Strategic and tactical price decisions in hotel revenue management.” *Tourism Management* 55: 123-132.

Anglin, P.M. and Gençay, R. 1996. "Semiparametric estimation of a hedonic price function." *Journal of Applied Econometrics* 11: 633-648.

Angrist, J.D. and Pischke, J.S. 2008. *Mostly harmless econometrics*. Princeton university press.

Anselin, L. and Arribas-Bel, D. 2013. “Spatial fixed effects and spatial dependence in a single cross-section.” *Papers in Regional Science* 92(1): 3-17.

Arvanitidis, P., Economou, A., Grigoriou, G. and Kollias, C. 2020. “Trust in peers or in the institution? A decomposition analysis of Airbnb listings' pricing.” *Current Issues in Tourism*, DOI: 10.1080/13683500.2020.1806794

Balaguer, J. and Pernías, J.C. 2013. “Relationship between spatial agglomeration and hotel prices.” Evidence from business and tourism consumers. *Tourism Management* 36: 391-400.

Baltagi, B.H., and Li, D. 2002. "Series estimation of partially linear panel data models with fixed effects." *Annals of Economics and Finance* 3: 103-116.

Becerra, M., Santaló, J., and Silva, R. 2013. “Being better vs. being different: Differentiation, competition, and pricing strategies in the Spanish hotel industry.” *Tourism Management* 34: 71-79.

Benítez-Aurioles, B. 2018. “Why are flexible booking policies priced negatively?” *Tourism Management* 67: 312-325.

Bishop, K.C. and Timmins, C. 2018. “Using panel data to easily estimate hedonic demand functions.” *Journal of the Association of Environmental and Resource Economists* 5(3): 517-543.

Boes, S. and Nüesch, S. 2011. “Quasi-experimental evidence on the effect of aircraft noise on apartment rents.” *Journal of Urban Economics* 69: 196-204.

Bontemps, C., Simioni, M. and Surry, Y. 2008. "Semiparametric hedonic price models: Assessing the effects of agricultural nonpoint source pollution." *Journal of Applied Econometrics* 23: 825-842.

Boto-García, D., 2022. “Multiway clustering in tourism research.” *Current Issues in Tourism* 25(3): 363-378.

Boto-García, D., Mayor, M. and DelaVega, P. 2021. “Spatial price mimicking on Airbnb: Multi-host vs single-host.” *Tourism Management* 87: 104365.

Breusch, T.S. and Pagan, A.R. 1980. “The Lagrange multiplier test and its applications to model specification in econometrics.” *Review of Economic Studies* 47: 239-253.

Cai, Y., Zhou, Y., Ma, J.J. and Scott, N. 2019. “Price determinants of Airbnb listings: Evidence from Hong Kong.” *Tourism Analysis* 24(2): 227-242.

Cameron, A.C. and Trivedi, P.K. 2009. *Microeconometrics using Stata*. College Station. Stata press.

Casamatta, G., Giannoni, S., Brunstein, D. and Jouve, J. 2022. “Host type and pricing on Airbnb: Seasonality and perceived market power.” *Tourism Management* 88: 104433.

Chattopadhyay, M. and Mitra, S.K. 2019. “Do Airbnb host listing attributes influence room pricing homogeneously?” *International Journal of Hospitality Management* 81: 54-64.

Chen, C. and Rothschild, R. 2010. “An application of hedonic pricing analysis to the case of hotel rooms in Taipei.” *Tourism Economics* 16(3): 685-694.

Chica-Olmo, J., González-Morales, J.G. and Zafra-Gómez, J.L. 2020. “Effects of location on Airbnb apartment pricing in Málaga.” *Tourism Management* 77: 103981.

Cinelli, C. and Hazlett, C. 2019. “Making sense of sensitivity: extending omitted variable bias.” *Journal of the Royal Statistical Society B: Methodological* 82(1): 39-67.

Cropper, M.L., Deck, L.B. and McConnell, K.E. 1988. "On the choice of functional form for hedonic price functions." *The Review of Economics and Statistics* 70(4): 668-675.

de Oliveira Santos, G.E. 2016. “Worldwide hedonic prices of subjective characteristics of hostels.” *Tourism Management* 52: 451-454.

Deboosere, R., Kerrigan, D.J., Wachmuth, D. and El-Geneidy, A. 2019. “Location, location and professionalization: a multilevel hedonic analysis of Airbnb listing prices and revenue.” *Regional Studies, Regional Science* 6(1): 143-156.

Ert, E., Fleischer, A. and Magen, N. 2016. 2Trust and reputation in the sharing economy: The role of personal photos in Airbnb.” *Tourism Management* 55: 62-73.

Espinet, J.M., Saez, M., Coenders, G. and Fluviá, M. 2003. “Effect on prices of the attributes of holiday hotels: a hedonic prices approach.” *Tourism Economics* 9(2): 165-177.

Falk, M., Larpin, B. and Scaglione, M. 2019. “The role of specific attributes in determining prices of Airbnb listings in rural and urban locations.” *International Journal of Hospitality Management* 83: 132-140.

Faye, B. 2021. “Methodological discussion of Airbnb's hedonic study. A review of the problems and some proposals tested on Bordeaux City data.” *Annals of Tourism Research* 86: 103079.

Fleischer, A. 2012. “A room with a view - A valuation of the Mediterranean Sea view.” *Tourism Management* 33: 598-602.

Frank, K.A. 2000. “Impact of a confounding variable on a regression coefficient.” *Sociological Methods & Research* 29(2): 147-194.

Gibbs, C., Guttentag, D., Gretzel, U., Morton, J. and Goodwill, A. 2018. “Pricing in the sharing economy: a hedonic pricing model applied to Airbnb listings.” *Journal of Travel & Tourism Marketing* 35(1): 46-56.

Greenstone, M. and Gallagher, J. 2008. “Does hazardous waste matter? Evidence from the housing market and the superfund program.” *The Quarterly Journal of Economics* 123(3): 951-1003.

Guttentag, D. 2015. “Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector.” *Current Issues in Tourism* 18(12): 1192-1217.

Gyodi, K. and Nawaro, L. 2021. “Determinants of Airbnb prices in European cities: A spatial econometrics approach.” *Tourism Management* 86: 104319.

Halvorsen, R. and Pollakowski, H.O. 1981. "Choice of functional form for hedonic price equations." *Journal of Urban Economics* 10: 37-49.

Hausman, J.A. 1978. “Specification tests in econometrics.” *Econometrica* 46(6): 1251-1271.

Hausman, J.A. and Taylor, W.E. 1981. “Panel data and unobservable individual effects.” *Econometrica* 49(6): 1377-1398.

Hill, R.J. 2013. "Hedonic price indexes for residential housing: A survey evaluation and taxonomy." *Journal of Economic Surveys* 27(5): 879-914.

IBESTAT 2021. *Institut d'Estadistica de les Illes Balears. Flujo de turistas extranjeros, FRONTUR*. Available at: https://ibestat.caib.es/ibestat/estadistiques/economia/turisme/0b70b294-81e0-413a-b7b2-3cc3a33593a8

Lancaster, K.J. 1966. “A new approach to consumer theory.” *Journal of Political Economy* 74(2): 132-157.

Latinopoulos, D. 2018. “Using a spatial hedonic analysis to evaluate the effect of sea view on hotel prices.” *Tourism Management* 65: 87-99.

Lawani, A., Reed, M.R., Mark, T. and Zheng, Y. 2019. “Reviews and price on online platforms: Evidence from sentiment analysis of Airbnb reviews in Boston.” *Regional Science and Urban Economics* 75: 22-34.

Leoni, V. 2020. “Stars vs lemons. Survival analysis of peer-to peer marketplaces: the case of Airbnb.” *Tourism Management* 79: 104091.

Linden, L. and Rockoff, J.E. 2008. “Estimates of the impact of crime risk on property values from Megan's Laws.” *American Economic Review* 98(3): 1103-1127.

Masiero, L., Nicolau, J.L. and Law, R. 2015. “A demand-driven analysis of tourist accommodation price: A quantile regression of room bookings.” *International Journal of Hospitality Management* 50: 1-8.

Moreno-Izquierdo, L., Ramón-Rodríguez, A.B., Such-Devesa, M.J. and Perles-Ribes, J.F. 2019. “Tourist environment and online reputation as a generator of added value in the sharing economy: The case of Airbnb in urban and sun-and-beach holiday destinations.” *Journal of Destination Marketing & Management* 11: 53-66.

Moreno-Izquierdo, L., Rubia-Serrano, A., Perles-Ribes, J.F., Ramón-Rodríguez, A.B. and Such-Devesa, M.J. 2020. “Determining factors in the choice of prices of tourist rental accommodation. New evidence using the quantile regression approach.” *Tourism Management Perspectives* 33: 100632.

Moulton, B. 1987. “Diagnostics for group effects in regression analysis.” *Journal of Business & Economic Statistics* 5(2): 275-282.

Mundlak, Y. 1978. “On the pooling of time series and cross section data.” *Econometrica* 46(1): 69-85.

Önder, I., Weismayer, C. and Gunter, U. 2019. “Spatial price dependencies between the traditional accommodation sector and the sharing economy.” *Tourism Economics* 25(8): 1150-1166.

Oster, E. 2019. “Unobservable selection and coefficient stability: Theory and evidence.” *Journal of Business & Economic Statistics* 37(2): 187-204.

Rigall-i-Torrent, R. and Fluvià, M. 2007. “Public goods in tourism municipalities: formal analysis, empirical evidence and implications for sustainable development.” *Tourism Economics* 13(3): 361-378.

Rigall-i-Torrent, R. and Fluvià, M. 2011. “Managing tourism products and destinations embedding public good components: A hedonic approach.” *Tourism Management* 32: 244-255.

Rigall-i-Torrent, R., Fluvià, M., Ballester, R., Saló, A., Ariza, E. and Espinet, J.M. 2011. "The effects of beach characteristics and location with respect to hotel prices." *Tourism Management* 32: 1150-1158.

Robinson, P.M. 1988. "Root-n-consistent semiparametric regression." *Econometrica* 56: 931-954.

Rosen, S. 1974. "Hedonic Prices and Implicit Markets: Product differentiation in pure competition." *Journal of Political Economy* 82(1): 34-55.

Sainagui, R., Abrate, G. and Mauri, A. 2021. "Price and RevPAR determinants of Airbnb listings: Convergent and divergent evidence." *International Journal of Hospitality Management* 92: 102709.

Saló, A., Garriga, A., Rigall-i-Torrent, R., Vila, M. and Fluvià, M. 2014. "Do implicit prices for hotels and second homes show differences in tourists' valuation for public attributes for each type of accommodation facility?" *International Journal of Hospitality Management* 36: 120-129.

Sánchez-Lozano, G., Nobre-Pereira, L. and Chávez-Miranda, E. 2021. "Big data hedonic pricing: Econometric insights into room rates' determinants by hotel category." *Tourism Management* 85: 104308.

Soler, I.P., Gemar, G., Correia, M.B. and Serra, F. 2019. "Algarve hotel price determinants: A hedonic pricing model." *Tourism Management* 70: 311-321.

Tang, L., Kim, J. and Wang, X. 2019. "Estimating spatial effects on peer-to-peer accommodation prices: Towards an innovative hedonic model approach." *International Journal of Hospitality Management* 81: 43-53.

Voltes-Dorta, A. and Sánchez-Medina, A. 2020. "Drivers of Airbnb prices according to property/room type, season and location: A regression approach." *Journal of Hospitality and Tourism Management* 45: 266-275.

Voltes-Dorta, A. and Inchausti-Sintes, F. 2021. "The spatial and quality dimensions of Airbnb markets." *Tourism Economics* 27(4): 688-702.

Wang, D. and Nicolau, J.L. 2017. "Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb.com." *International Journal of Hospitality Management* 62: 120-131.

Wang, Y. and Blei, D.M. 2019. "The blessings of multiple causes." *Journal of the American Statistical Association* 114(528): 1574-1596.

Xie, K. and Mao, Z. 2019. "Locational strategy for professional hosts: Effect on perceived quality and revenue performance of Airbnb listings." *Journal of Hospitality & Tourism Research* 43(6): 919-929.

Yang, Y., Mao, Z. and Tang, J. 2018. "Understanding guest satisfaction with urban hotel location." *Journal of Travel Research* 57: 243-259.

Zhang, H., Zhang., J., Cheng, S. and Zhang, J. 2011. "Modeling hotel room price with geographically weighted regression." *International Journal of Hospitality Management* 30: 1036-1043.

Table 1. Summary statistics

Variable	Mean/ %	Std. Dev.	Min	Max
Dep. Variable: Price (€)	352.26	421.55	11	7,029
Intrinsic characteristics				
Apartment	42.40			
House	37.04			
Villa	15.82			
Other property	4.72			
Entire	84.86			
Shared/private	15.13			
Bedrooms	2.59	1.68	0	10
Bathrooms	2.17	1.43	0	8
Minimum Stay	3.91	2.21	1	21
Num. Photos	25.99	21.44	1	469
Business ready	5.49			
Host characteristics				
Superhost	7.52			
Num. listings	4.53	6.82	1	37
Rental policies				
Instant booking	10.88			
Cancellation policy: Flexible	6.33			
Cancellation policy: Moderate	7.60			
Cancellation policy: Strict	86.06			
Competitive rivalry				
Num. competitors	263.59	217.75	2	654
Num. hotel beds	9689.55	5707.77	464	16369
Num. reservations past month	7.64	9.34	1	31
Temporal factors				
August 2015	9.71			
September 2015	10.55			
October 2015	8.08			
November 2015	4.00			
December 2015	3.40			
January 2016	3.71			
February 2016	3.39			
March 2016	4.51			
April 2016	6.06			
May 2016	7.52			
June 2016	7.48			
July 2016	9.93			
August 2016	10.00			
September 2016	9.59			
Number of listings	1,990			
Number of time periods	14			
Number of observations	11,939			

Table 2. Hedonic regression parameter estimates

Dependent variable: ln Price	(1)	(2)	(3)	(4)	(5)
Explanatory variables	Pooled OLS	RE	FE	Mundlak	Taylor-Hausman
Time invariant					
Apartment	-0.293***	-0.203**		-0.199**	-0.244
	(0.093)	(0.103)		(0.082)	(0.173)
House	-0.048	0.004		0.008	0.060
	(0.073)	(0.086)		(0.077)	(0.162)
Villa	0.050	0.094		0.102	0.192
	(0.080)	(0.090)		(0.083)	(0.178)
Entire	0.795***	0.731***		0.740***	0.713***
	(0.037)	(0.032)		(0.035)	(0.074)
Bedrooms	0.172***	0.193***		0.168***	0.185***
	(0.020)	(0.020)		(0.017)	(0.026)
Bathrooms	0.205***	0.201***		0.198***	0.193***
	(0.025)	(0.023)		(0.019)	(0.033)
Minimum stay	0.011	0.019***		0.010*	0.020**
	(0.007)	(0.006)		(0.006)	(0.008)
Num. photos	-3.4e-04	1.2e-04		0.001	4.8e-04
	(0.001)	(0.001)		(0.001)	(0.001)
Business ready	0.059	0.060		0.013	0.024
	(0.045)	(0.043)		(0.047)	(0.053)
Superhost	0.083*	0.030		0.091**	-2.451***
	(0.045)	(0.032)		(0.038)	(0.621)
Instant book	-0.080***	-0.040		-0.039	-0.133*
	(0.025)	(0.031)		(0.032)	(0.074)
Canc. Policy: flexible	0.087**	0.121**		0.080*	0.065
	(0.040)	(0.048)		(0.043)	(0.077)
Can. Policy: moderate	-0.056*	-0.065**		-0.045	0.013
	(0.032)	(0.029)		(0.031)	(0.072)
Time variant					
Num. listings	4.0e-04	2.7e-04	-0.014	-0.014	-0.013
	(0.003)	(0.002)	(0.010)	(0.010)	(0.010)
Num. competitors	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***
	(4.4e-04)	(3.0e-03)	(1.7e-04)	(2.8e-04)	(0.000)
Hotel beds	-3.7e-06	-9.3e-07	-8.4e-08	-8.4e-07	-8.3e-07
	(7.2e-06)	(5.6e-06)	(1.5e-06)	(5.4e-06)	(5.5e-06)
Num. reservations past month	-0.013***	-0.001*	-2.4e-04	-2.4e-04	-2.9e-04
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
September 2015	-0.092***	-0.041***	-0.036***	-0.036***	-0.036***
	(0.015)	(0.008)	(0.008)	(0.008)	(0.008)
October 2015	-0.372***	-0.238***	-0.226***	-0.226***	-0.227***
	(0.057)	(0.030)	(0.030)	(0.030)	(0.030)
November 2015	-0.512***	-0.372***	-0.361***	-0.361***	-0.361***
	(0.111)	(0.075)	(0.074)	(0.074)	(0.075)
December 2015	-0.463***	-0.342***	-0.334***	-0.334***	-0.333***
	(0.114)	(0.074)	(0.073)	(0.073)	(0.073)
January 2016	-0.412***	-0.289***	-0.281***	-0.281***	-0.281***
	(0.117)	(0.076)	(0.075)	(0.075)	(0.075)
February 2016	-0.460***	-0.346***	-0.339***	-0.339***	-0.338***
	(0.112)	(0.079)	(0.078)	(0.078)	(0.078)
March 2016	-0.486***	-0.333***	-0.324***	-0.324***	-0.323***
	(0.122)	(0.081)	(0.079)	(0.079)	(0.079)
April 2016	-0.434***	-0.323***	-0.316***	-0.316***	-0.316***
	(0.119)	(0.074)	(0.071)	(0.071)	(0.072)
May 2016	-0.269***	-0.225***	-0.223***	-0.223***	-0.223***
	(0.055)	(0.036)	(0.035)	(0.035)	(0.035)
June 2016	-0.078**	-0.045**	-0.044**	-0.044**	-0.044**

	(0.035)	(0.018)	(0.017)	(0.017)	(0.017)
July 2016	0.160***	0.158***	0.156***	0.156***	0.156***
	(0.021)	(0.013)	(0.013)	(0.013)	(0.013)
August 2016	0.267***	0.254***	0.251***	0.251***	0.252***
	(0.015)	(0.020)	(0.020)	(0.020)	(0.020)
September 2016	-0.036	0.009	0.010	0.010	0.010
	(0.026)	(0.009)	(0.009)	(0.009)	(0.009)
Mean Num. listings				0.014	
				(0.011)	
Mean Hotel beds				-1.0e-05	
				(2.0e-05)	
Mean Num. competitors				0.003**	
				(0.001)	
Mean Num. reservations past month				-0.023***	
				(0.002)	
Means of monthly dummies	NO	NO	NO	YES	NO
Postal code dummies	YES	YES	YES	YES	YES
Constant	4.154***	3.872***	5.873***	4.119***	4.171***
	(0.104)	(0.121)	(0.059)	(0.352)	(0.170)
Observations	11,939	11,939	11,939	11,939	11,939
Number of time periods	14	14	14	14	14
Number of IDs	1,990	1,990	1,990	1,990	1,990

Clustered standard errors at the zip code level in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3. Sensitivity measures to uncontrolled quality based on Cinelli and Hazlett (2019)

Explanatory variables	$RV(\%)$	$RV_{\alpha=0.05}\ (\%)$	${R^2}_{Y\sim w\|W}(\%)$
Time invariant			
Apartment	11.52	9.92	1.48
House	1.93	0.15	0.04
Villa	1.82	0.04	0.03
Entire	38.13	37.04	19.02
Bedrooms	24.68	23.33	7.48
Bathrooms	27.11	25.81	9.16
Minimum stay	4.53	2.80	0.21
Num. photos	1.42	-	0.02
Business ready	2.73	0.96	0.08
Superhost	4.35	2.61	0.20
Instant book	4.90	3.17	0.25
Canc. Policy: flexible	4.21	2.47	0.18
Can. Policy: moderate	2.99	1.23	0.09
Time variant			
Num. listings	0.53	-	0.01
Num. competitors	3.73	1.99	0.14
Hotel beds	1.09	-	0.01
Num. reservations past month	19.53	18.08	4.53

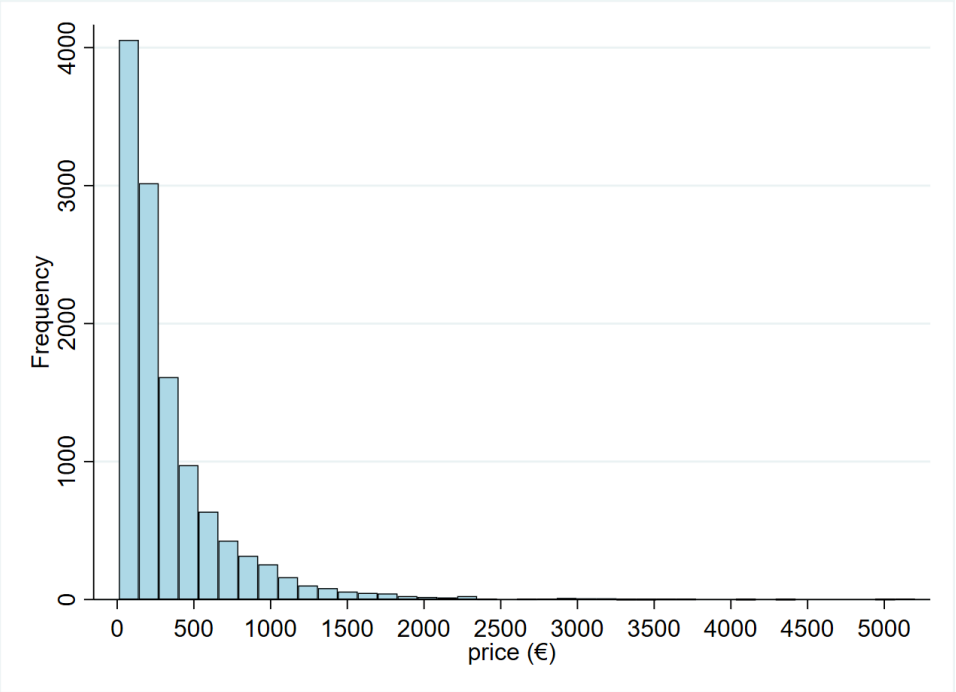

Figure 1. Histogram of average daily price

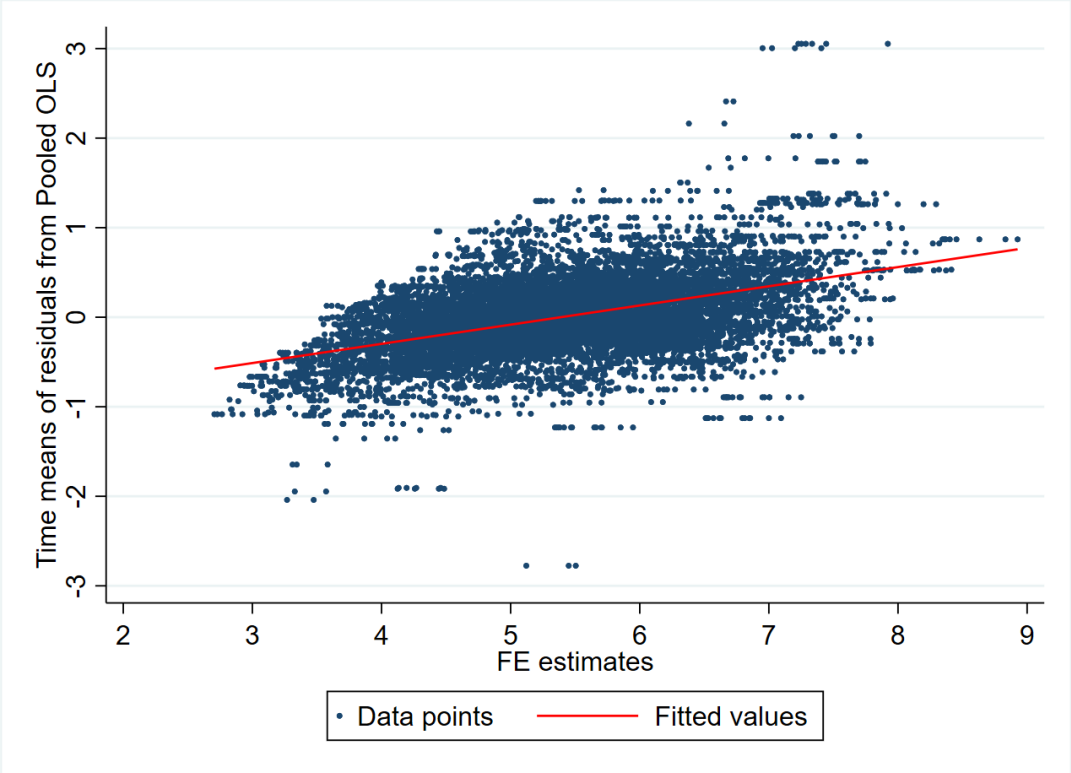

Figure 2. Scatterplot of time means of residuals from Pooled OLS against fixed effects estimates

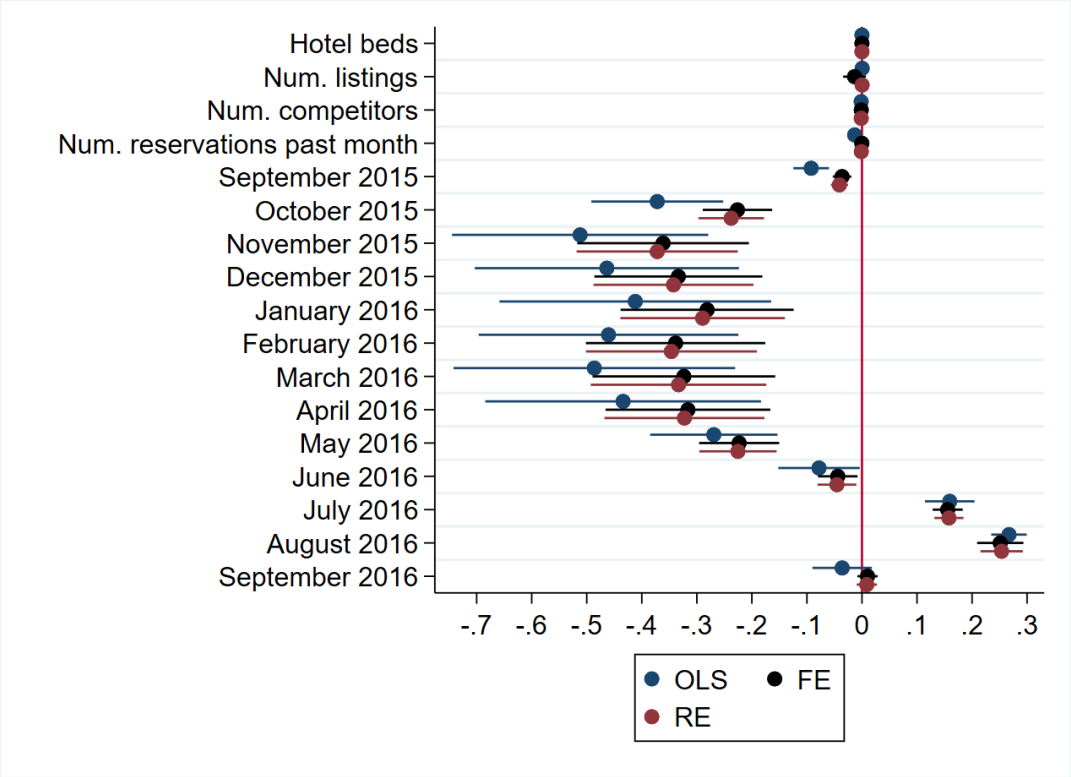

Figure 3. Coefficient estimates for the time-variant variables obtained from OLS, FE and RE

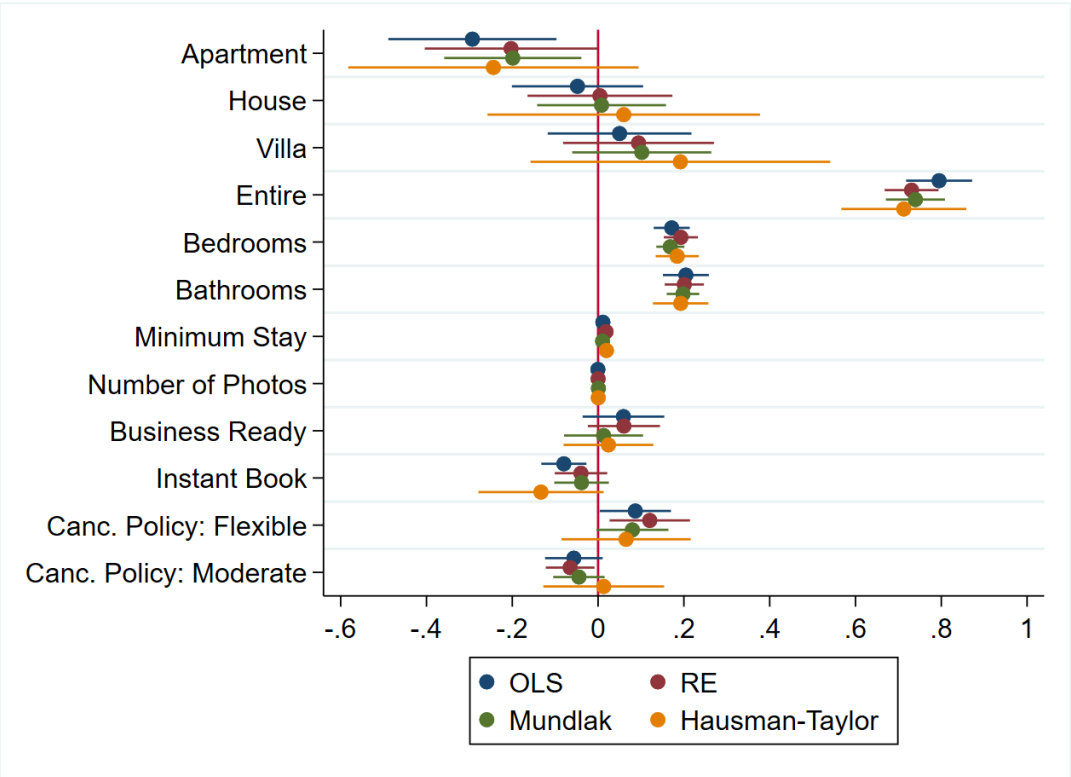

Figure 4. Coefficient estimates for the time-invariant variables (except Superhost) obtained from OLS, RE, Mundlak and Taylor-Hausman

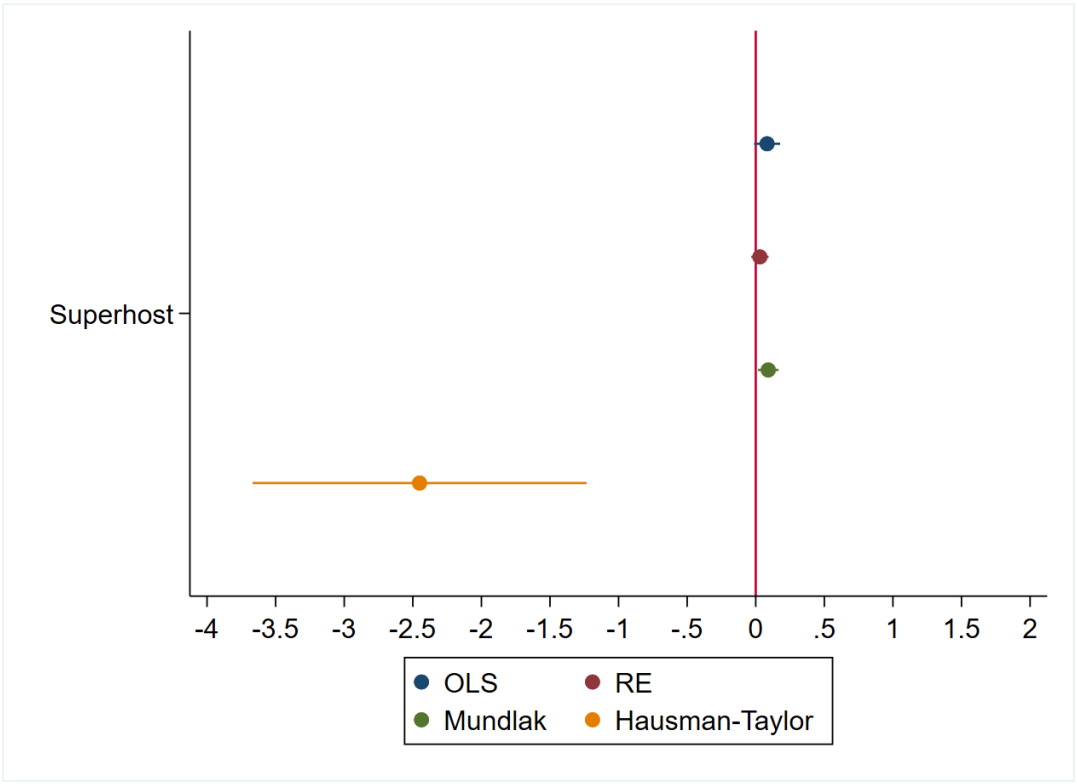

Figure 5. Coefficient estimates for Superhost obtained from OLS, RE, Mundlak and Taylor-Hausman

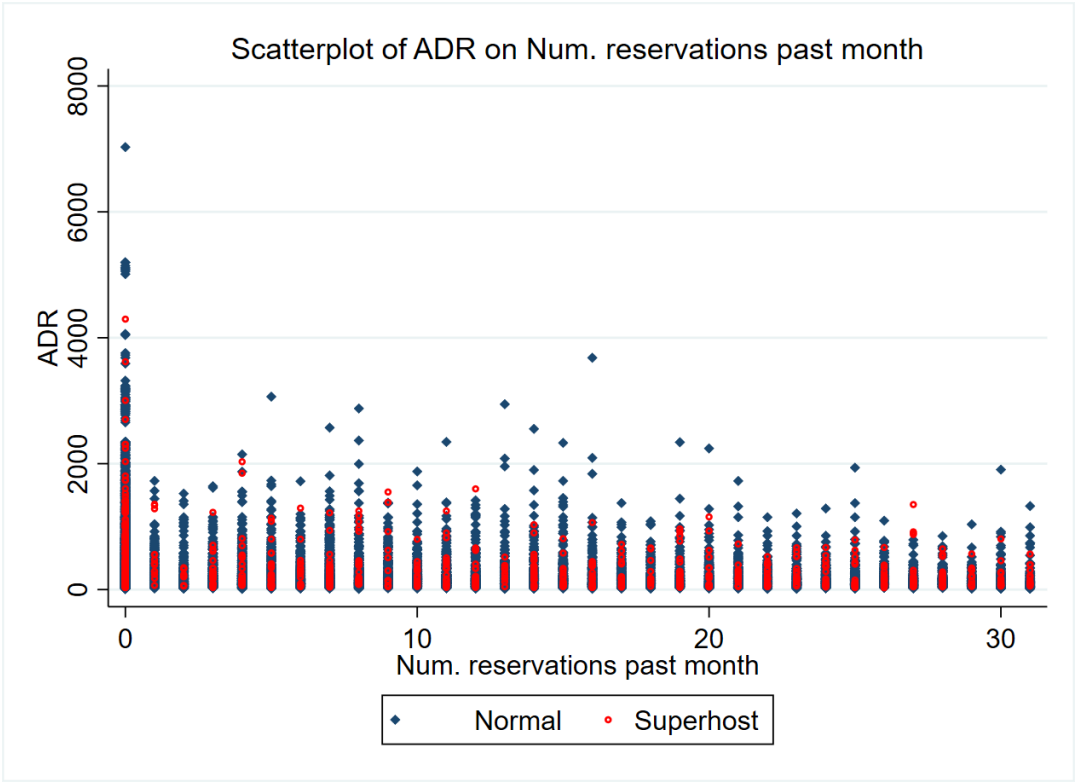

Figure 6. Scatterplot of ADR on Num. reservations past month

ENDNOTES

[1] Importantly, our analysis assumes linear-in-parameters hedonic price functions. Semiparametric and machine learning approaches are beyond the scope of the paper.

[2] We focus on the bias stemming from omitted attributes that are constant over time. Omitted factors that are property-specific and vary over time could also be an aspect of concern, although they are less common in empirical hedonic studies. Most of the quality dimensions of an accommodation property are time-invariant.

[3] The *Superhost* badge is a quality indicator that is conceded by the platform to hosts that meet the following criteria: (i) completed a minimum of 10 stays that sum up to 100 nights; (ii) maintained a response rate of 90% or higher; (iii) maintained a cancellation rate of 1% or less; and (v) maintained a general rate of 4.8/5 in the last 365 days (Airbnb, 2021).

[4] We assume potential guests have superior information about home attributes than the analyst. This assumption seems plausible, since they can inspect pictures, have a look at reviews or directly ask the host for specific information. By contrast, researchers usually work with datasets with a limited number of listing characteristics. Moreover, even if the research has data about the reviews or can access the photographs, their qualitative nature hinders considering them into the regression framework.

[5] All listings j that belong to the same geographical area k (country, city, district, postal code, neighbourhood, etc.) are assumed to share the same locational price premium. The relevant unobserved ambient factors are assumed to be defined within the boundaries of the geographical disaggregation unit considered.

[6] These methods assume linear parametric specifications. The recent *deconfounder* method proposed by Wang and Blei (2019) is a valuable alternative for detecting omitted variable bias in non-linear outcome models.

[7] Algebraically, a pooled OLS estimator applied to (5) in principle would produce similar estimates. Nevertheless, the induced correlation in the composed error term though μ_j makes feasible GLS more appropriate.

[8] This dataset has been previously used by Leoni (2020) in a study of listings' survivability in the Airbnb marketplace.

[9] This requires properties to compile with several requirements like having a self-check-in solution, a WiFi connection, a laptop-friendly workspace (a private desk or table) or a high rating (over 4.8), among others.

[10] Nonetheless, for robustness, the analysis is also done using the Box-Cox transformation $\widetilde{Price} = (Price^{0.034} - 1)/0.034$. As presented in Figure A1 and Table A2 in Supplementary Material, because the estimated lambda is close to zero (log-transformation), the estimation results are very close to the ones presented in the main analysis.

[11] The average Variance Inflation Factor (VIF) after OLS is 2.17. Therefore, the model specification does not suffer from collinearity problems.

[12] As mentioned before, the methods by Oster (2019) and Frank (2000) are valid alternatives.

[13] Please notice the values of $RV_{\alpha=0.05}$ (%) are missing from Table 3 for Num. photos, Num. listings and Hotel beds. The reason is that these three variables are not found to be significant in the OLS regressions and have point estimates that are virtually zero. That is why it is unfeasible to compute an indicator of how large confounders should be to make the point estimate statistically zero (since it is already).

[14] This is further confirmed by auxiliary checks using propensity score matching showing that the *Superhost* indicator explains a substantial part of the listings' fixed effects (available upon request).