



UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS

DÉPARTEMENT DES SCIENCES ADMINISTRATIVES

WORKING PAPER No. 4

Hedonic Pricing of Short-Term Rental Properties: A Multi-Method Analysis of Airbnb Listing Prices in Quebec

Simon-Pierre Boucher

Département des sciences administratives

Université du Québec en Outaouais

simon-pierre.boucher@uqo.ca

Gatineau – Pavillon Alexandre-Taché

283, boulevard Alexandre-Taché

Gatineau, Québec, Canada J9A 1L8

October 2026

Version 1.1

Abstract

We estimate hedonic price functions for a province-wide cross-section of Airbnb listings in Quebec, Canada, from a scrape of 5,000 listing cards collected in early 2026 (3,456 with a displayed price). The displayed price field is bimodal: about 28% of priced listings, mostly small Montreal rental units, show the total price of a stay of roughly three weeks rather than a nightly rate. A two-regime mixture of hedonic regressions estimated by EM separates the two populations sharply and leaves 2,480 nightly prices, on which all estimation is based. A semi-logarithmic hedonic model with city fixed effects, estimated with heteroskedasticity-robust and clustered standard errors, explains 51% of the variance of log prices. Guest capacity (elasticity 0.33), bathrooms (0.44) and bedrooms (0.16), a shared bathroom (−32%), the number of photos, the displayed rating (3.4% per tenth of a point) and local listing density are the priced attributes; the Superhost and Guest Favourite badges carry no premium conditional on capacity, type and location. Quantile regressions show that cheap listings are priced per guest and certified by reviews, whereas expensive listings are priced per room and are cheaper when heavily reviewed. Spatial lag, error and Durbin models find moderate spatial dependence ($\hat{\rho} \approx 0.20$, $\hat{\lambda} \approx 0.28$) with direct impacts within a few percent of the OLS coefficients. Gradient-boosting models raise the out-of-sample R^2 from 0.47 to 0.56 under random cross-validation and from 0.41 to 0.48 when whole areas are held out, and their SHAP decomposition ranks the same attributes as the hedonic model. The paper provides the first provincial hedonic evidence for Quebec’s short-term rental market and a template for detecting and handling stay-price contamination in scraped platform prices.

Keywords: hedonic price function, short-term rentals, Airbnb, implicit prices, quantile regression, spatial econometrics, machine learning, Quebec

JEL Classification: C21, C31, C45, L83, R31, Z30

1. Introduction

A short-term rental listing is a bundle of characteristics priced by an individual host rather than by a centralised market maker. The theory of hedonic prices ([Rosen, 1974](#)), which builds on the characteristics view of demand of [Lancaster \(1966\)](#), treats the observed price of such a differentiated good as the sum of the implicit prices of its attributes, and the hedonic price function has become the standard empirical tool for housing and property markets ([Malpezzi, 2003](#); [Sirmans et al., 2005](#)). Short-term rental platforms offer an unusually complete view of this function: every listing displays its capacity, its type, its amenities, its host's status and its reputation, together with a price that the host chose. Since [Wang and Nicolau \(2017\)](#) and [Gibbs et al. \(2018a\)](#), a sizeable literature has estimated hedonic price equations for Airbnb listings in individual cities and across cities ([Chen and Xie, 2017](#); [Teubner et al., 2017](#); [Pérez-Sánchez et al., 2018](#); [Cai et al., 2019](#); [Lorde et al., 2019](#); [Gyódi and Nawaro, 2021](#); [Tong and Gunter, 2022](#)), with the recurrent finding that capacity, room type, location and reputation signals explain most of the cross-sectional variation in displayed prices, and with a secondary literature on the value of trust signals such as Superhost status, review scores and host photos ([Ert et al., 2016](#); [Zervas et al., 2021](#)).

Quebec is an interesting and little-studied case. The province combines a large metropolitan market (Montreal), a heritage tourism market (Quebec City) and an extensive resort and cottage market in the Laurentians, Charlevoix and the Eastern Townships, so that a provincial sample mixes urban apartments with rural chalets in a way that single-city studies cannot. Canadian evidence on short-term rentals has concentrated on their housing-market consequences ([Wachsmuth and Weisler, 2018](#); [Combs et al., 2020](#)); the only hedonic study we are aware of, [Deboosere et al. \(2019\)](#), estimates a multilevel model of prices and revenue for Montreal and finds that accessibility and professionalisation dominate. No provincial hedonic analysis exists, and none uses the newer platform signals (the Guest Favourite badge, introduced in 2023) or compares the hedonic model with quantile, spatial and machine-learning estimators on the same data.

This paper estimates hedonic price functions for a province-wide cross-section of Airbnb listings in Quebec scraped from the public search pages of the platform in early 2026. The data are a snapshot of 5,000 listing cards, of which 3,456 display a price. They are not an Inside Airbnb extract: the record is the listing card, which carries the title, the displayed price, the star rating, the review count, capacity, bedroom and bathroom counts, an amenity count, the Superhost and Guest Favourite badges, the pet policy and the coordinates, but neither the room type, the host identifier, the amenity list, the cleaning fee nor the minimum stay. The limited field set is a constraint that shapes the paper, and we describe exactly what each variable measures rather than relabelling it.

The first task of the paper is a measurement problem that any user of scraped Airbnb prices must confront. The displayed price field is bimodal on the logarithmic scale: one mode lies near \$150

and a second near \$2,500, and 26.5% of priced listings carry a “nightly” price above \$1,000. The second mode is populated by ordinary one- and two-bedroom rental units in Montreal, its listings have prices that are about twenty times those of otherwise similar listings in the first mode, and the pattern is best explained by the platform displaying the total price of a multi-night stay for part of the search results. We document this with a two-component mixture model and a regime regression, and we estimate the hedonic models on the subsample priced at or below \$1,000, with the threshold varied in robustness checks. The companion paper on Airbnb and rents (UQO Working Paper No. 5) uses the same scrape and reaches the same diagnosis; here we treat it as the point of departure.

The second task is the hedonic analysis itself. We estimate a sequence of nested ordinary least squares specifications that add amenity, host and reputation, location and finally city fixed effects to a structural baseline, with heteroskedasticity-robust and city-clustered standard errors; quantile regressions at five quantiles of the conditional log-price distribution, with bootstrap tests of inter-quantile differences; spatial lag, spatial error and spatial Durbin models estimated by maximum likelihood on a nearest-neighbour weights matrix, with direct and indirect impacts computed as in [LeSage and Pace \(2009\)](#); and five supervised learning models evaluated under both random and spatially blocked cross-validation ([Roberts et al., 2017](#)), with the best model interpreted through SHAP values ([Lundberg and Lee, 2017](#); [Lundberg et al., 2020](#)). Throughout, the estimates are read as conditional associations in a cross-section of asking prices, not as structural marginal willingness to pay ([Kuminoff et al., 2010](#); [Bishop et al., 2020](#)).

Three results stand out. First, once the stay-price regime is removed, the hedonic model behaves as the literature leads one to expect: capacity variables, title category and location explain the bulk of price variation, with an R^2 that rises from about one-third in the structural baseline to about one-half with city fixed effects, compared with R^2 values of 0.05 to 0.27 obtained on the contaminated full sample in the first version of this paper. Second, reputation signals carry precisely estimated but modest premia once capacity and location are held fixed: Superhost status and the Guest Favourite badge are each worth a few percent of the nightly price, the review score matters mainly at the lower quantiles, and the number of reviews is negatively associated with price, consistent with high-turnover listings being cheaper. Third, flexible estimators improve on the fixed-effects hedonic model by a meaningful but bounded margin: gradient-boosting models raise the out-of-sample R^2 from 0.47 to 0.56 under random cross-validation and from 0.41 to 0.48 when whole areas are held out, every model loses about seven points of R^2 in the second design, and the SHAP decomposition of the best model ranks capacity first and the badges last, as the hedonic coefficients do, while exposing one sign-switching association (the review count) that a constant coefficient averages away. The non-linearities that the learners capture refine the hedonic surface rather than overturn it.

The contribution is therefore twofold. Substantively, the paper provides the first province-wide hedonic estimates for Quebec’s short-term rental market and documents how implicit prices differ

between the urban and the resort segments and across the price distribution. Methodologically, it shows how a measurement defect of scraped platform prices can be detected and handled, and it benchmarks, on one data set, how much the choice of estimator matters relative to the choice of sample.

The remainder of the paper is organised as follows. Section 2 describes the data, the price-field diagnosis and the analysis sample. Section 3 sets out the hedonic, quantile, spatial and machine-learning specifications and the inference used. Section 4 reports the results, Section 5 the robustness checks, and Section 6 concludes with limitations and directions for future work.

2. Data

2.1. Source and Content of the Scrape

The data are a scrape of Airbnb’s public search pages for the province of Quebec collected in early 2026. The file carries no timestamp; it was assembled in the same period as the January 2026 rental extraction used in the companion paper on Airbnb and rents (UQO Working Paper No. 5), which analyses the same 5,000 listings. The unit of observation is the listing card as it appears in search results, and the record contains nineteen fields: the card title, the municipality, a property-type label, the displayed price, a price category, the star rating and a rating category, the numbers of bedrooms and bathrooms, the guest capacity, the number of amenities, the number of photos, the number of reviews, a composite “quality score” generated by the scraper, the Superhost and Guest Favourite badges, the pet policy and the coordinates. The data are therefore considerably sparser than an Inside Airbnb extract: the room type (entire home, private room, shared room), the host identifier, the amenity list, the cleaning fee, the minimum stay and the calendar are not observed, and the first version of this paper, which described the data as an Inside Airbnb sample with amenity lists and host profiles, was wrong on this point.¹

The card title follows a fixed template, for example “Rental unit in Montréal · ★4.90 · 2 bedrooms · 3 beds · 1 shared bath”, and it carries information that the structured fields do not. We parse it to obtain the title category assigned by the host (44 distinct labels, which we group into six classes), the star rating as displayed, a “New” marker for listings that have no rating yet, the number of beds, and whether the bathroom is shared, private or a half bath. Parsing also repairs three defects of the structured fields. First, the rating field is missing for 456 of the 3,456 priced listings (13.2%) and had been median-imputed in the committed version of the data; the title shows that these listings display

¹The scraper-generated quality score is a deterministic function of the review count, the badges and the photo count ($R^2 = 0.63$ in a linear regression on those fields) and is not used. Three fields of the first version are no longer used: “amenity indices” that were computed by adding random noise to the normalised amenity count, “room-type” indicators that relabelled the property-type field, and a “multi-listing host” proxy based on duplicated card titles.

no star, so we treat them as unrated and add an indicator. Second, the scraper stored zero bathrooms whenever the title qualified the bathroom (“1 shared bath”, “1 private bath”, “Half-bath”); the count is recovered from the title for 465 listings. Third, 161 studios have zero bedrooms in the structured field and are recoded as one sleeping room with a studio indicator, and 46 bedroom counts missing from the field are read from the title. Fifty titles were served in German by the platform’s locale and are parsed with the corresponding patterns; for 37 of them the rating field had been truncated to 4.0 by a decimal-comma parsing error and the title value is retained.

2.2. The Price Field

Of the 5,000 cards, 1,544 (30.9%) display no price and are dropped. The displayed price of the remaining 3,456 listings has a median of \$220 and a mean of \$1,233, and its distribution on the logarithmic scale is bimodal (Figure 1). A two-component normal mixture fitted to $\ln p$ places one mode near \$150 and the other near \$2,500; 26.5% of listings (915) display a price above \$1,000. The upper mode is not a luxury segment. Table 1 shows that, bedroom count by bedroom count, the median price of listings above \$1,000 is sixteen to nineteen times the median of listings below it, a ratio that no plausible quality difference produces but that a stay of about three weeks does. Listings in the upper mode are disproportionately small rental units in Montreal with few reviews and without the Superhost badge, which is the profile of listings oriented towards stays of a month or more under Montreal’s short-term rental rules. We conclude that the scraper captured, for part of the sample, the total price of a multi-night stay rather than the nightly rate, and the companion paper reaches the same diagnosis on the same file. Because neither the stay length nor the search dates are recorded, the nightly price of these listings cannot be recovered.

Table 1. Median Displayed Price by Bedroom Count, Below and Above \$1,000

Bedrooms	<i>N</i>	Median (\$)	<i>N</i>	Median (\$)	Ratio
		≤ \$1,000		> \$1,000	
0	15	105	2	4,679	44.6
1	953	112	439	2,072	18.5
2	648	162	226	2,735	16.9
3	508	209	129	3,884	18.6
4	277	270	72	4,624	17.1
5+	140	366	47	3,665	10.0

Notes: Priced listings ($N = 3,456$). Studios are counted as one bedroom; the “0” row contains listings whose bedroom count could not be parsed. The ratio is the median above \$1,000 divided by the median below.

A single price threshold is an imperfect classifier: the three-week total of a \$40 room is below \$1,000, and the genuine nightly rate of a large chalet can exceed it. We therefore classify listings with

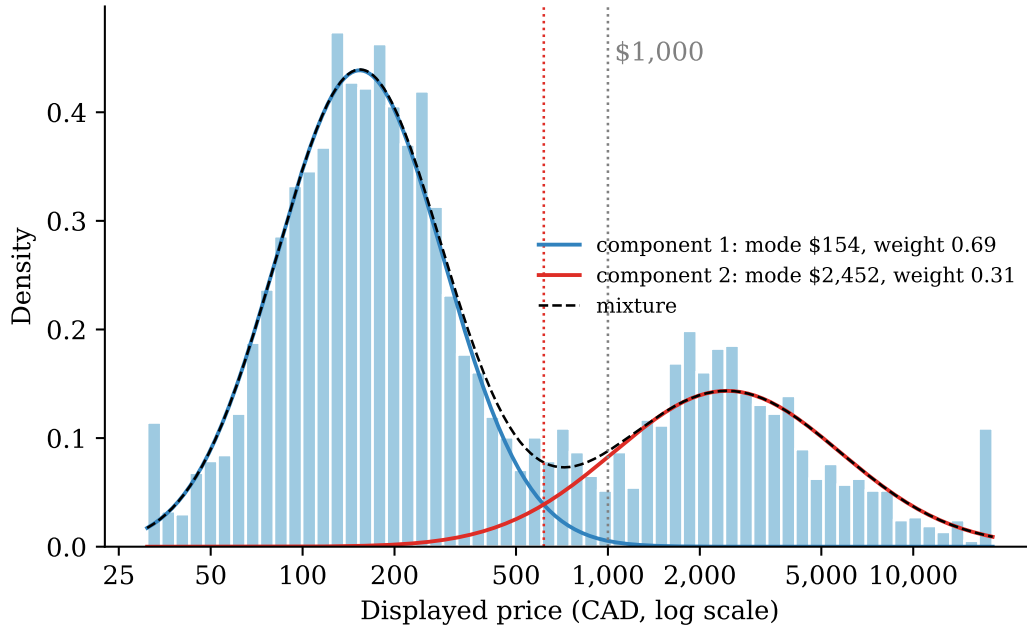


Figure 1. Displayed Prices of 3,456 Quebec Airbnb Listings on a Logarithmic Scale

Notes: Histogram of $\ln p$ with the two components of a normal mixture fitted by maximum likelihood (solid lines) and their sum (dashed). The grey dotted line marks \$1,000; the red dotted line marks the price at which the posterior probability of the upper component reaches one half.

the mixture of hedonic regressions of Section 3.6, which uses the listing's characteristics to judge whether its price is a nightly or a stay price. Table 2 reports the estimates. The stay regime shifts the intercept by $\hat{\delta} = 3.15$ log points, an implied stay of about 23 nights, and has a wider residual dispersion, as expected if stay lengths vary; its probability rises in Montreal and falls with capacity, review count, Superhost status and the hotel and chalet categories. The posterior classification is sharp: only 7 listings have a posterior probability between 0.25 and 0.75. It assigns 976 listings (28.2%) to the stay regime, of which 83 display a price at or below \$1,000, and it keeps 22 listings priced above \$1,000, mostly large chalets, in the nightly regime. The 2,480 listings with a posterior probability below one half form the analysis sample.

2.3. Analysis Sample and Descriptive Statistics

Table 3 describes the analysis sample. The median displayed price is \$162 and the mean \$206; the 10th and 90th percentiles are \$70 and \$376. The typical listing has two bedrooms, one and a half bathrooms and sleeps six; 91% display a rating, with a mean of 4.81 and a tenth percentile of 4.57 among rated listings, the compression documented by Zervas et al. (2021); 59% of hosts are Superhosts and 41% of listings carry the Guest Favourite badge. The sample spans 438 municipalities; Montreal accounts for 20% of listings, Quebec City and Mont-Tremblant for the next largest shares, and the median listing is 17 km from the nearest of the 22 reference city centres.

Table 2. Two-Regime Mixture of Hedonic Regressions: Nightly Versus Stay Prices

Parameter	Estimate
Intercept shift of the stay regime, δ (log points)	3.154
Implied stay length, $\exp(\delta)$ (nights)	23.4
Residual s.d., nightly regime, σ_0	0.493
Residual s.d., stay regime, σ_1	0.568
Mean posterior probability of the stay regime	0.283
Listings classified in the stay regime ($\hat{p}_i \geq 0.5$)	976
of which displayed price $\leq \$1,000$	83
Listings classified nightly with displayed price $> \$1,000$	22
Listings with $0.25 < \hat{p}_i < 0.75$	7
Log-likelihood	-4,323.4
EM iterations	16
<i>Regime equation (logit coefficients, SE)</i>	
Constant	1.466*** (0.141)
Montreal	0.622*** (0.096)
$\ln(1 + \text{reviews})$	-0.331*** (0.027)
$\ln(\text{guests})$	-0.889*** (0.073)
Superhost	-0.588*** (0.087)
Hotel/B&B/hostel	-0.947*** (0.191)
Chalet/cottage/cabin	-0.255** (0.121)

Notes: Estimates of Equation (8) by the EM algorithm on the 3,456 priced listings. The common slopes β use the M4 regressor set plus a Montreal indicator. The regime equation is a logistic regression of the posterior regime probability on the listed characteristics; standard errors in parentheses. ***, ** and * denote significance at the 1, 5 and 10% levels.

The last column shows that the full priced sample differs from the analysis sample mainly in price (because of the stay regime) and in the Montreal share. Prices in the analysis sample are winsorised at the 1st and 99th percentiles of the full priced sample (\$31 and \$18,205) by the committed cleaning step, which affects 35 listings at the bottom (clipped to \$31) and none at the top.

Table 3. Descriptive Statistics, Analysis Sample

Variable	Mean	Std. dev.	P10	Median	P90	<i>N</i>	Mean, full
Displayed price (CAD)	206	176	70	162	376	2,480	1,233
ln(price)	5.08	0.68	4.25	5.08	5.93	2,480	5.89
Bedrooms	2	1	1	2	4	2,480	2
Bathrooms	1.52	0.98	1.00	1.00	2.50	2,480	1.46
Guest capacity	6	3	2	6	10	2,480	5
Beds (from title)	4	3	1	3	6	2,443	3
Amenities count	48	15	26	48	67	2,480	46
Photos	31	20	11	26	56	2,480	28
Reviews	64	79	0	37	162	2,480	53
Rating (rated listings)	4.81	0.20	4.57	4.87	5.00	2,246	4.81
Rated (star shown)	0.91	0.29	1.00	1.00	1.00	2,480	0.87
Superhost	0.59	0.49	0.00	1.00	1.00	2,480	0.53
Guest Favourite	0.41	0.49	0.00	0.00	1.00	2,480	0.34
Pets allowed	0.30	0.46	0.00	0.00	1.00	2,480	0.29
Shared bathroom	0.06	0.23	0.00	0.00	0.00	2,480	0.07
Studio	0.03	0.18	0.00	0.00	0.00	2,480	0.05
Distance to nearest centre (km)	48.24	259.14	1.22	17.42	70.19	2,480	41.46
Listings within 1 km	24	45	0	3	94	2,480	25
Montreal	0.20	0.40	0.00	0.00	1.00	2,480	0.27
Posterior prob. of stay regime	0.00	0.02	0.00	0.00	0.00	2,480	0.28

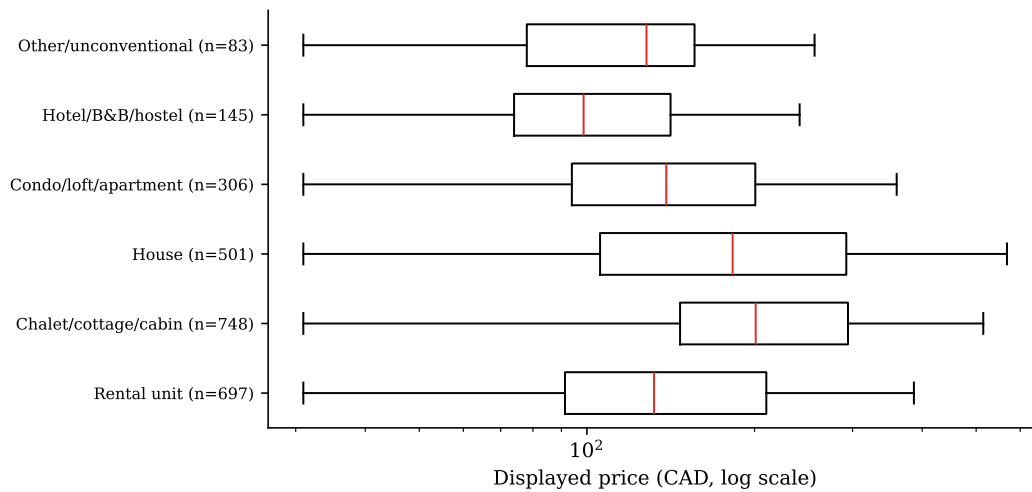
Notes: Analysis sample of 2,480 listings classified in the nightly-price regime. The last column reports the mean in the full priced sample ($N = 3,456$). Binary variables are 0/1. Bathrooms and bedrooms are the repaired counts described in Section 2.1. Distances are great-circle distances to the nearest of 22 reference municipalities.

Table 4 and Figure 2 show prices by title category. Chalets, cottages and cabins (30% of the sample) and houses (20%) have the highest median prices, about \$200 and \$180; rental units (28%) and condominiums, lofts and apartments (12%) are priced near \$135; hotel rooms, bed-and-breakfasts and hostels (6%) near \$100. These raw differences confound type with capacity and location, which the hedonic model separates. Figure 3 maps the analysis sample: the Montreal and Quebec City agglomerations, the Laurentian corridor north of Montreal, Charlevoix and the Eastern Townships are the main concentrations, and the density of listings within 1 km ranges from zero for a third of the sample to over 90 in central Montreal.

Table 4. Displayed Nightly Price by Title Category, Analysis Sample

Title category	N	Share	Mean (\$)	Median (\$)	P10	P90
Rental unit	697	0.281	167	132	67	290
Chalet/cottage/cabin	748	0.302	259	201	110	466
House	501	0.202	241	183	67	469
Condo/loft/apartment	306	0.123	166	139	65	277
Hotel/B&B/hostel	145	0.058	118	99	53	191
Other/unconventional	83	0.033	148	128	32	238
All	2,480	1.000	206	162	70	376

Notes: Title categories group the 44 labels hosts choose for the listing card. “Chalet/cottage/cabin” includes vacation and holiday homes and nature lodges; “House” includes townhouses, bungalows, villas and farm stays; “Condo/loft/apartment” includes serviced apartments and guest suites; “Hotel/B&B/hostel” includes apart-hotels, boutique hotels, guesthouses and resorts.

**Figure 2.** Displayed Nightly Price by Title Category

Notes: Box plots (median, interquartile range, whiskers at 1.5 times the interquartile range; outliers omitted) on a logarithmic price axis, analysis sample.

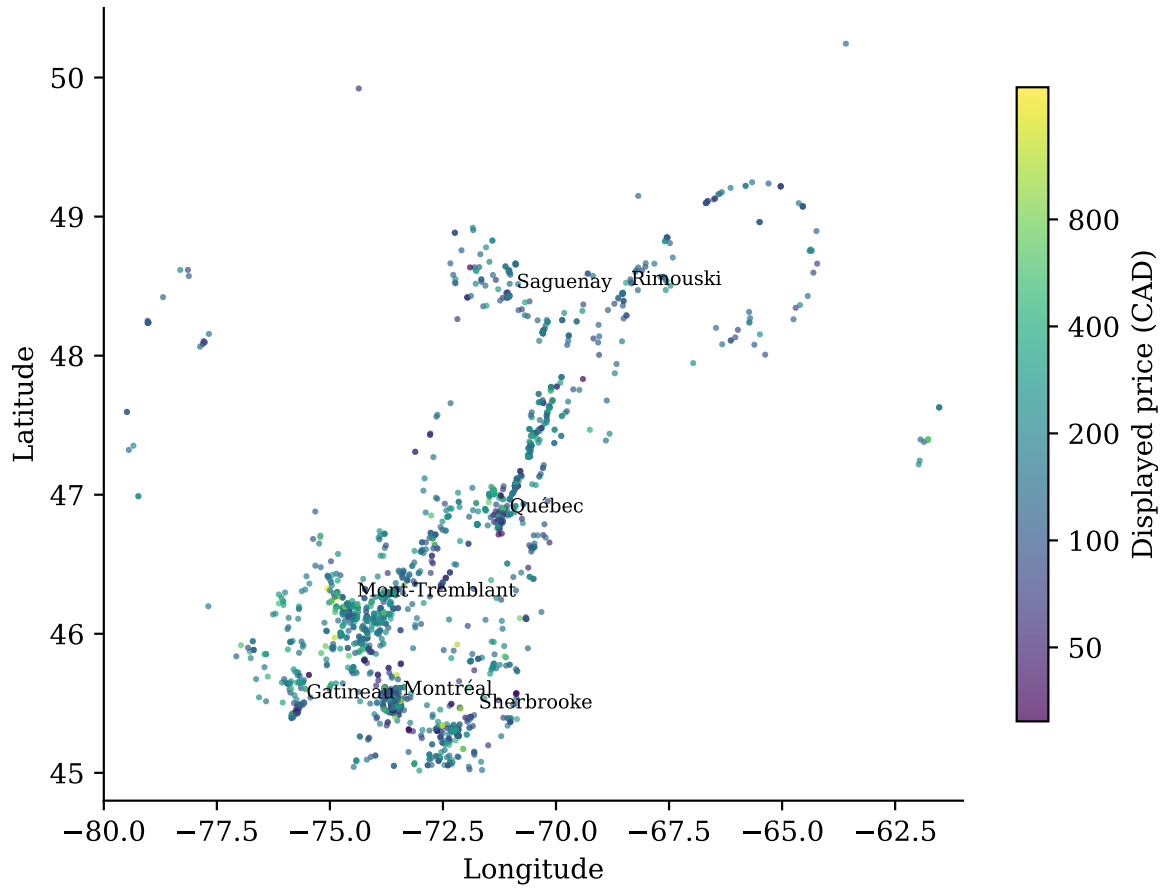


Figure 3. Geographic Distribution of the Analysis Sample

Notes: Each point is a listing in the nightly-price regime, coloured by displayed price on a logarithmic scale. Coordinates are WGS-84; a small number of listings in the Gaspésie and on the North Shore lie outside the plotted window.

2.4. What the Cross-Section Can and Cannot Measure

Three features of the data bound the interpretation of everything that follows. The price is the host's asking price displayed on one day, not a transaction price, and it embodies whatever dynamic pricing the host applies (Gibbs et al., 2018b); the cleaning fee, which is a substantial and variable component of what guests pay, is not observed. The sample is a single cross-section: the first version of this paper referred to "seasons" and "seasonal sub-markets", but no seasonal variation exists in the data, and seasonality can enter only through the composition of listings active at the time of the scrape, which the city fixed effects absorb. Finally, the scrape covers listings returned by the platform's search for Quebec on the collection dates, so listings that were calendar-blocked or filtered out are missing, and the province-wide coverage is denser in the urban and resort areas where the platform is most active.

3. Methodology

3.1. Hedonic Price Function

Let p_i denote the displayed nightly price of listing i and $\mathbf{x}_i$ the vector of its observed characteristics. Following Rosen (1974), the hedonic price function $P(\cdot)$ maps characteristics into prices, and we approximate it by the semi-logarithmic form that dominates the housing and short-term rental literatures (Malpezzi, 2003; Sirmans et al., 2005; Wang and Nicolau, 2017):

$$\ln p_i = \alpha + \beta'_S \mathbf{S}_i + \beta'_A \mathbf{A}_i + \beta'_H \mathbf{H}_i + \beta'_L \mathbf{L}_i + \mu_{c(i)} + \varepsilon_i, \quad (1)$$

where $\mathbf{S}_i$ collects the structural variables ($\ln$ bedrooms, $\ln(1 + \text{bathrooms})$, $\ln$ guest capacity, a shared-bathroom indicator, a studio indicator and five title-category indicators with *Rental unit* as the reference), $\mathbf{A}_i$ the listing-level presentation variables ($\ln(1 + \text{amenities})$, $\ln(1 + \text{photos})$, pets allowed), $\mathbf{H}_i$ the host and reputation variables (Superhost, Guest Favourite, $\ln(1 + \text{reviews})$, a rated indicator and the centred rating interacted with it), $\mathbf{L}_i$ the location variables ($\ln(1 + \text{km to the nearest centre})$ and $\ln(1 + \text{listings within 1 km})$), and $\mu_{c(i)}$ a fixed effect for the city or spatial cell $c(i)$ in which the listing is located. The error ε_i is mean zero conditional on the regressors but is allowed to be heteroskedastic and spatially correlated within cities.

The semi-logarithmic form has three consequences for interpretation. For a continuous regressor entered in logarithms the coefficient is an elasticity. For a continuous regressor in levels, such as the rating, 100β is the approximate percentage change in price per unit. For an indicator variable the percentage premium is

$$g = 100 \left[\exp\left(\hat{\beta} - \frac{1}{2}\widehat{\text{Var}}(\hat{\beta})\right) - 1 \right], \quad (2)$$

the [Kennedy \(1981\)](#) correction of the [Halvorsen and Palmquist \(1980\)](#) transformation, which removes the small-sample bias of $\exp(\hat{\beta}) - 1$ and is the form we report in the text. The coefficients in Equation (1) are conditional associations between characteristics and asking prices in a cross-section; they are not marginal willingness to pay, which would require the stronger assumptions discussed by [Kuminoff et al. \(2010\)](#) and [Bishop et al. \(2020\)](#), nor realised transaction prices.

3.2. Specifications and Inference

We estimate Equation (1) by ordinary least squares in five nested specifications: M1 contains the structural block only; M2 adds the presentation block; M3 the host and reputation block; M4 the two location variables; and M5 the city fixed effects. Cities with at least 15 listings in the analysis sample receive their own effect; the remaining listings, which are scattered across several hundred small municipalities, are grouped into 25 spatial cells obtained by k -means clustering of their projected coordinates, so that every listing belongs to exactly one fixed-effect group. The fixed effects absorb all between-city variation in prices, including tourism demand, seasonality of the local market and regulation, and the remaining coefficients are identified from within-city variation.

Standard errors are heteroskedasticity-robust (HC3, [MacKinnon and White, 1985](#)) in all specifications. Because the regressors of neighbouring listings are correlated and the error term is likely to be spatially correlated, we also report standard errors clustered by fixed-effect group for M5 ([Cameron and Miller, 2015](#)), and we re-estimate M5 with 60 finer spatial cells as the fixed-effect unit in the robustness section. Multicollinearity among the capacity variables is monitored through variance inflation factors, and block significance is assessed with robust F tests.

3.3. Quantile Regression

To examine whether implicit prices differ between budget and premium listings we estimate the conditional quantile function of [Koenker and Bassett \(1978\)](#),

$$Q_{\tau}(\ln p_i \mid \mathbf{x}_i) = \alpha_{\tau} + \beta'_{\tau} \mathbf{x}_i + \mu_{c(i),\tau}, \quad \tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}, \quad (3)$$

with the full M5 regressor set, by minimising the check-function loss $\sum_i \rho_{\tau}(\ln p_i - \alpha_{\tau} - \beta'_{\tau} \mathbf{x}_i - \mu_{c(i),\tau})$, $\rho_{\tau}(u) = u(\tau - \mathbf{1}\{u < 0\})$ ([Koenker and Hallock, 2001](#)). Standard errors use the kernel estimate of the sparsity function. Goodness of fit is the Koenker–Machado pseudo- R^2 , one minus the ratio of the minimised loss to the loss of the unconditional quantile. Because the coefficients at different quantiles are estimated from the same sample, differences between the $\tau = 0.90$ and $\tau = 0.10$ coefficients are tested with a paired bootstrap of 100 replications in which both quantile regressions are re-estimated on each resample.

3.4. Spatial Econometric Models

Listings cluster in space and their prices may depend on neighbours' prices or on spatially correlated unobservables. We construct a row-standardised k -nearest-neighbour weights matrix $\mathbf{W}$ ($k = 5$, with $k = 10$ as a check) from coordinates projected to kilometres, so that the anisotropy of degrees of longitude at Quebec's latitude does not distort distances, and we test the OLS residuals of M4 and M5 for spatial autocorrelation with Moran's I ,

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} \hat{u}_i \hat{u}_j}{\sum_i \hat{u}_i^2}, \quad (4)$$

with inference from 999 random permutations. We then estimate by maximum likelihood (Anselin, 1988; LeSage and Pace, 2009) the spatial lag model (SAR), the spatial error model (SEM) and the spatial Durbin model (SDM),

$$\text{SAR: } \ln \mathbf{p} = \rho \mathbf{W} \ln \mathbf{p} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (5)$$

$$\text{SEM: } \ln \mathbf{p} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} = \lambda \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}, \quad (6)$$

$$\text{SDM: } \ln \mathbf{p} = \rho \mathbf{W} \ln \mathbf{p} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad (7)$$

using the M4 regressor set, since the city fixed effects and the spatial lag compete for the same between-city variation. The SDM nests the SAR ($\boldsymbol{\theta} = \mathbf{0}$) and the SEM ($\boldsymbol{\theta} = -\rho\boldsymbol{\beta}$), and we use a likelihood-ratio test of the first restriction (Elhorst, 2010).

In models with a spatial lag the coefficients are not marginal effects. Writing $\mathbf{S}_k(\mathbf{W}) = (\mathbf{I} - \rho\mathbf{W})^{-1}(\mathbf{I}\boldsymbol{\beta}_k + \mathbf{W}\boldsymbol{\theta}_k)$ for regressor k (with $\boldsymbol{\theta}_k = \mathbf{0}$ in the SAR), the average direct effect is $N^{-1} \text{tr} \mathbf{S}_k(\mathbf{W})$, the average total effect is $N^{-1} \boldsymbol{\iota}' \mathbf{S}_k(\mathbf{W}) \boldsymbol{\iota}$, and the indirect (spillover) effect is their difference (LeSage and Pace, 2009). We compute the three summary measures exactly from the dense inverse and obtain their standard errors from 300 draws of the parameters from their estimated asymptotic distribution.

3.5. Machine Learning Benchmarks and Interpretation

The log-linear form imposes additive, constant implicit prices. To measure how much predictive accuracy this restriction costs we compare the M5 specification with an elastic net (Zou and Hastie, 2005), a random forest (Breiman, 2001), gradient boosting (Friedman, 2001), XGBoost (Chen and Guestrin, 2016) and LightGBM, all trained on the same regressors, including the fixed-effect indicators. Performance is measured out of sample by five-fold cross-validation in two designs: random folds, which is the usual benchmark, and spatially blocked folds in which the 20 spatial clusters of listings are assigned to folds as whole blocks (Roberts et al., 2017). The second design

asks whether a model trained on some areas predicts prices in areas it has never seen, which is the relevant question for an automated valuation tool and a stricter test of generalisation when observations are spatially dependent. We report the out-of-sample R^2 , root mean squared error and mean absolute error on the log scale and the median absolute percentage error (MdAPE) on the price scale, median $_i |\exp(\hat{y}_i) - p_i|/p_i$. The first version of this paper reported a mean absolute percentage error computed on log prices, which is not interpretable as a price error and is not used here.

The best learner is interpreted with SHAP values (Lundberg and Lee, 2017), computed exactly for tree ensembles (Lundberg et al., 2020). For a prediction $\hat{f}(\mathbf{x}_i)$ the SHAP values ϕ_{ik} satisfy $\hat{f}(\mathbf{x}_i) = \phi_0 + \sum_k \phi_{ik}$ with $\phi_0 = \mathbb{E}[\hat{f}]$, and the mean absolute value $\overline{|\phi_k|} = N^{-1} \sum_i |\phi_{ik}|$ ranks features by their average contribution to predicted log prices. SHAP values describe the fitted model, not the data-generating process; they are attribution weights for associations captured by the learner and carry no causal content.

3.6. Identifying the Nightly-Price Regime

Section 2 shows that the displayed price field mixes nightly prices with total prices of multi-night stays. A price threshold separates the two regimes imperfectly because the stay total of a cheap listing can fall below the threshold while the nightly price of a large chalet can exceed it. We therefore model the regime explicitly with a two-component mixture of hedonic regressions with common slopes,

$$\ln p_i = \mathbf{x}_i' \boldsymbol{\beta} + \delta z_i + \varepsilon_i, \quad \varepsilon_i \mid z_i \sim \mathcal{N}(0, \sigma_{z_i}^2), \quad \Pr(z_i = 1 \mid \mathbf{w}_i) = \Lambda(\mathbf{w}_i' \boldsymbol{\gamma}), \quad (8)$$

where $z_i = 1$ indicates the stay regime, δ is the log of the typical stay length that the displayed price covers, the regime-specific variance allows stay lengths to vary across listings, and the regime probability depends through a logistic link on characteristics $\mathbf{w}_i$ (Montreal location, review count, capacity, Superhost status and the hotel and chalet categories) that the descriptive analysis identifies as predictors of the second mode. The model is estimated by the EM algorithm, initialised at the \$1,000 rule; the E-step computes the posterior probability $\hat{p}_i = \Pr(z_i = 1 \mid \ln p_i, \mathbf{x}_i, \mathbf{w}_i)$ and the M-step updates $(\boldsymbol{\beta}, \delta)$ by weighted least squares on the stacked data, (σ_0, σ_1) from the weighted residuals and $\boldsymbol{\gamma}$ by a weighted logistic regression. Listings with $\hat{p}_i < 0.5$ constitute the analysis sample used in all subsequent estimation; the robustness section varies the cut-off, re-weights by $1 - \hat{p}_i$ and compares with the simple price-threshold rules.

4. Empirical Results

4.1. Nested Hedonic Models

Table 5 reports the five nested specifications. The structural block alone (M1) explains 44% of the variance of log prices in the analysis sample, a figure in line with single-city Airbnb studies (Wang and Nicolau, 2017; Chen and Xie, 2017; Gyódi and Nawaro, 2021) and far above the 5% obtained on the contaminated price field. Adding the presentation variables (M2) and the host and reputation variables (M3) raises R^2 by about one percentage point each; the two location variables (M4) add two points, and the city and cell fixed effects (M5) a further three, to 0.509 (adjusted 0.495). The fixed effects alone explain 13% of the variance, so the structural and listing-level characteristics carry most of the explanatory power and location acts mainly through the composition of listings across places rather than through large city-level price differences. The root mean squared error of M5 is 0.475 log points, which is a typical prediction error of about 40% of the nightly price for an individual listing.

The coefficients are stable across specifications, which indicates that the omitted-variable bias from leaving out reputation or location is small for the structural variables. Clustering the standard errors by fixed-effect group raises them by 20 to 60% for the capacity and location variables and leaves the inference qualitatively unchanged, except that the bedroom elasticity falls from the 1% to the 5% level and the rating coefficient becomes more precise. Variance inflation factors in M4 peak at 4.2 (for ln bedrooms), so the capacity variables, although correlated, are separately identified (Appendix Table A.3).

4.2. Capacity, Type and Presentation

Capacity is priced through three channels. Holding bedrooms and bathrooms constant, a 10% increase in guest capacity is associated with a 3.3% higher price (elasticity 0.331, HC3 standard error 0.037); holding capacity constant, a 10% increase in one plus the bathroom count is associated with a 4.4% higher price (0.437, 0.070), and a 10% increase in bedrooms with a 1.6% higher price (0.160, 0.042). The bedroom elasticity is modest because capacity and bathrooms absorb most of the size effect; in M1 without the other capacity variables it would be far larger. A shared bathroom, which in these data identifies the room-in-a-home segment that the missing room-type field would otherwise have captured, carries a premium of $g = 100[\exp(-0.385 - \frac{1}{2}0.053^2) - 1] = -32\%$ by Equation (2), the largest single implicit price in the model. Studios are priced 16% above an otherwise equivalent one-bedroom listing of the same capacity, which reflects their urban location and the fact that a studio of given capacity is a complete dwelling rather than a room.

Title categories matter less than their raw price differences in Table 4 suggest. Without location

Table 5. Nested Hedonic Specifications, Dependent Variable $\ln(\text{Displayed Nightly Price})$

	(1)	(2)	(3)	(4)	(5)	(5')
Dependent variable: $\ln(\text{price})$	M1	M2	M3	M4	M5	M5
$\ln(\text{bedrooms})$	0.168*** (0.043)	0.154*** (0.042)	0.158*** (0.042)	0.164*** (0.042)	0.160*** (0.042)	0.160** [0.068]
$\ln(1 + \text{bathrooms})$	0.560*** (0.063)	0.514*** (0.064)	0.507*** (0.065)	0.464*** (0.067)	0.437*** (0.070)	0.437*** [0.089]
$\ln(\text{guests})$	0.351*** (0.038)	0.324*** (0.037)	0.327*** (0.038)	0.321*** (0.037)	0.331*** (0.037)	0.331*** [0.035]
Shared bathroom	-0.441*** (0.053)	-0.429*** (0.052)	-0.423*** (0.053)	-0.403*** (0.052)	-0.385*** (0.053)	-0.385*** [0.061]
Studio	0.189*** (0.061)	0.210*** (0.063)	0.213*** (0.063)	0.179*** (0.062)	0.147** (0.062)	0.147* [0.089]
Chalet/cottage/cabin	0.045 (0.028)	0.010 (0.029)	0.003 (0.029)	0.154*** (0.033)	0.099*** (0.037)	0.099** [0.039]
House	-0.028 (0.031)	-0.035 (0.031)	-0.037 (0.032)	0.095*** (0.033)	0.079** (0.034)	0.079 [0.049]
Condo/loft/apartment	-0.052 (0.037)	-0.063* (0.036)	-0.071* (0.036)	-0.043 (0.035)	-0.032 (0.037)	-0.032 [0.038]
Hotel/B&B/hostel	0.008 (0.046)	0.024 (0.047)	0.044 (0.047)	0.122** (0.049)	0.079 (0.051)	0.079* [0.045]
Other/unconventional	0.042 (0.080)	0.063 (0.080)	0.047 (0.081)	0.188** (0.092)	0.135 (0.094)	0.135 [0.102]
$\ln(1 + \text{amenities})$		0.071** (0.035)	0.048 (0.037)	0.067* (0.037)	0.094** (0.038)	0.094** [0.038]
Pets allowed		-0.048** (0.023)	-0.044* (0.023)	-0.032 (0.023)	-0.044* (0.023)	-0.044 [0.029]
$\ln(1 + \text{photos})$		0.115*** (0.022)	0.104*** (0.022)	0.092*** (0.022)	0.087*** (0.022)	0.087*** [0.024]
Superhost			0.005 (0.024)	0.008 (0.024)	0.023 (0.025)	0.023 [0.032]
Guest Favourite			-0.055** (0.027)	-0.034 (0.027)	-0.035 (0.027)	-0.035 [0.032]
$\ln(1 + \text{reviews})$			0.020** (0.010)	0.015 (0.010)	0.015 (0.010)	0.015 [0.026]
Rated (star shown)			-0.056 (0.049)	-0.033 (0.050)	-0.041 (0.050)	-0.041 [0.044]
Rating – 4.8 (rated only)			0.310*** (0.069)	0.328*** (0.068)	0.337*** (0.068)	0.337*** [0.065]
$\ln(1 + \text{km to centre})$				-0.002 (0.010)	-0.035* (0.019)	-0.035 [0.024]
$\ln(1 + \text{listings within 1 km})$				0.067*** (0.009)	0.064*** (0.013)	0.064*** [0.023]
City / cell fixed effects	No	No	No	No	Yes	Yes
Standard errors	HC3	HC3	HC3	HC3	HC3	Cluster
Observations	2,480	2,480	2,480	2,480	2,480	2,480
R^2	0.443	0.454	0.461	0.479	0.509	0.509
Adjusted R^2	0.440	0.451	0.457	0.475	0.495	0.495
RMSE	0.506	0.501	0.498	0.490	0.475	0.475
AIC	3,686	3,639	3,619	3,539	3,490	3,490

Notes: OLS on the analysis sample ($N = 2,480$). Columns (1)–(5) report HC3 standard errors in parentheses; column (5') repeats M5 with standard errors clustered by fixed-effect group (51 groups) in brackets. The reference title category is *Rental unit*. Rating is centred at 4.8 and set to zero for unrated listings, so the “Rated” indicator is the premium of a listing rated 4.8 over an unrated one. City / cell fixed effects: one indicator per municipality with at least 15 listings (26 cities) and 25 k -means cells for the remaining listings. ***, ** and * denote significance at the 1, 5 and 10% levels.

controls (M1–M3) no category differs significantly from rental units; once distance and density are held fixed (M4) and especially within cities (M5), chalets, cottages and cabins are priced 10% above rental units of the same capacity (Kennedy-corrected premium 10.3%, $p < 0.01$) and houses 8% above ($p < 0.05$), while condominiums and lofts, hotel rooms and unconventional listings are not significantly different. The reversal between M3 and M4 arises because chalets are located where listings are sparse and far from city centres, two attributes that are themselves priced; the within-city chalet premium is the better estimate of the type's implicit price.

Among the presentation variables the number of photos is the most robust: a 10% increase in one plus the photo count is associated with a 0.9% higher price in every specification. The amenity count has a smaller and less stable coefficient (elasticity 0.09 with fixed effects, $p < 0.05$), and allowing pets is associated with a price about 4% lower, significant at the 10% level. The first version of this paper reported large negative “amenity index” effects; those indices were noise and the amenity count, which is the only amenity information in the data, is positively priced.

4.3. Reputation Signals

The reputation block adds little to the fit, and its coefficients are small relative to the capacity effects. The displayed rating is the one reputation variable with a sizeable and precisely estimated coefficient: among rated listings, each additional point of rating is associated with a 34% higher price (0.337, standard error 0.068 in M5), so that the typical spread between a listing rated 4.6 and one rated 5.0 corresponds to about 14%. Because ratings are compressed near the top of the scale, with a tenth percentile of 4.57 (the pattern documented by [Zervas et al., 2021](#)), this coefficient is identified from small differences and should be read as the price distance between the lower tail and the mass of near-perfect ratings, not as a linear effect across the whole scale. Unrated listings, mostly new ones, are priced 4% below rated listings at 4.8, a difference that is not significant.

Superhost status and the Guest Favourite badge carry no premium once capacity, type and location are controlled: the Superhost coefficient is 0.023 (standard error 0.025) and the Guest Favourite coefficient -0.035 (0.027). Both badges are strongly associated with price in the raw data, but in the direction of the stay-price regime (Appendix Table A.2), not of the nightly price. This null is consistent with the mixed evidence in the literature, where [Teubner et al. \(2017\)](#) and [Gibbs et al. \(2018a\)](#) find Superhost premia of a few percent in some cities and none in others, and with the view that the badge rewards responsiveness and reliability rather than quality that guests pay for. The number of reviews is likewise unrelated to price on average (0.015, 0.010); the quantile results below show that this average conceals opposite signs at the two ends of the distribution. Figure 4 displays the partial relationship between the rating and log price and between distance to the nearest centre and log price as binned partial residuals from M4.

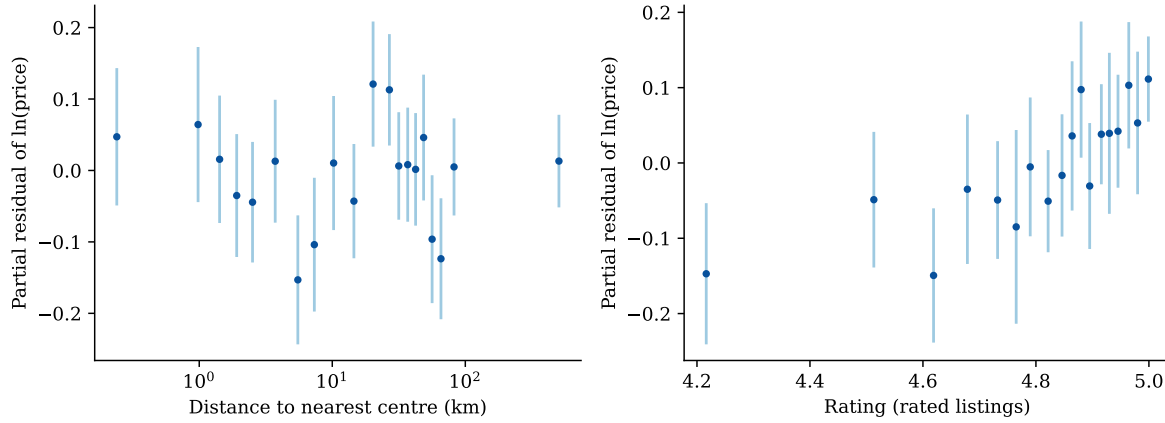


Figure 4. Partial Relationships: Distance to Centre and Rating

Notes: Binned means (20 quantile bins) of the partial residual $\hat{u}_i + \hat{\beta}_k x_{ik}$ from specification M4, with 95% confidence intervals. Left: distance to the nearest of 22 reference centres (symmetric-log axis). Right: displayed rating, rated listings only.

4.4. Location

The two location variables behave as the urban-economics literature predicts once the sample is restricted to nightly prices. Local density is positively priced: a 10% increase in one plus the number of listings within 1 km is associated with a 0.6% higher price (0.064, 0.013 with fixed effects), the opposite of the “competitive pressure” reading in the first version of this paper and consistent with density proxying for centrality and amenity within cities. Distance to the nearest reference centre has a small negative coefficient that is significant only at the 10% level with fixed effects (−0.035, 0.019); across the province, distance mixes the within-city gradient with the fact that resort listings are far from any city centre and expensive, which the category and density variables partly absorb. The fixed effects themselves are jointly significant (robust F test, $p < 0.001$), and the robustness section shows that finer spatial cells leave the listing-level coefficients essentially unchanged.

4.5. Quantile Regression

Table 6 and Figure 5 report the quantile estimates. The pseudo- R^2 rises from 0.278 at $\tau = 0.10$ to 0.405 at $\tau = 0.90$: observed characteristics explain more of the variation among expensive listings than among cheap ones, the reverse of what the first version reported on the contaminated field.

Three patterns emerge. First, the implicit prices of bedrooms and bathrooms increase along the distribution: the bedroom elasticity is indistinguishable from zero at $\tau = 0.10$ and 0.19 at $\tau = 0.90$, and the bathroom elasticity doubles from 0.24 to 0.55, while the capacity elasticity moves the other way, from 0.37 to 0.31. Cheap listings are priced per guest; expensive listings are priced per room. The paired bootstrap puts the $\tau = 0.90$ minus $\tau = 0.10$ difference in the bedroom elasticity at 0.17 (standard error 0.10, $p = 0.08$) and finds no significant difference for capacity (−0.07, $p = 0.37$);

Table 6. Quantile Regression Estimates, Specification M5

Variable	OLS	$\tau = 0.10$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.90$
ln(bedrooms)	0.160*** (0.042)	0.014 (0.053)	0.099** (0.039)	0.169*** (0.033)	0.214*** (0.032)	0.187*** (0.037)
ln(1 + bathrooms)	0.437*** (0.070)	0.241*** (0.078)	0.298*** (0.053)	0.401*** (0.044)	0.551*** (0.047)	0.551*** (0.058)
ln(guests)	0.331*** (0.037)	0.374*** (0.048)	0.399*** (0.036)	0.346*** (0.031)	0.305*** (0.031)	0.307*** (0.037)
Shared bathroom	-0.385*** (0.053)	-0.267*** (0.067)	-0.332*** (0.051)	-0.414*** (0.045)	-0.388*** (0.047)	-0.309*** (0.056)
Chalet/cottage/cabin	0.099*** (0.037)	0.024 (0.053)	0.068* (0.040)	0.117*** (0.036)	0.166*** (0.036)	0.214*** (0.043)
House	0.079** (0.034)	0.017 (0.048)	0.040 (0.037)	0.099*** (0.033)	0.087*** (0.033)	0.121*** (0.039)
Condo/loft/apartment	-0.032 (0.037)	-0.183*** (0.049)	-0.032 (0.037)	-0.017 (0.034)	0.014 (0.034)	0.077** (0.039)
Hotel/B&B/hostel	0.079 (0.051)	-0.091 (0.069)	0.101* (0.053)	0.181*** (0.047)	0.107** (0.047)	0.063 (0.058)
ln(1 + amenities)	0.094** (0.038)	0.053 (0.048)	0.110*** (0.036)	0.086*** (0.033)	0.032 (0.034)	0.017 (0.042)
ln(1 + photos)	0.087*** (0.022)	0.085*** (0.028)	0.100*** (0.022)	0.104*** (0.019)	0.100*** (0.021)	0.133*** (0.027)
Superhost	0.023 (0.025)	0.081** (0.032)	0.052** (0.025)	0.006 (0.023)	0.025 (0.023)	-0.003 (0.028)
Guest Favourite	-0.035 (0.027)	-0.050 (0.036)	-0.034 (0.028)	0.000 (0.025)	-0.024 (0.025)	-0.080*** (0.029)
ln(1 + reviews)	0.015 (0.010)	0.042*** (0.015)	0.037*** (0.010)	0.007 (0.008)	-0.026*** (0.008)	-0.042*** (0.008)
Rated (star shown)	-0.041 (0.050)	-0.052 (0.064)	-0.056 (0.048)	0.054 (0.041)	0.056 (0.041)	0.072 (0.047)
Rating – 4.8 (rated only)	0.337*** (0.068)	0.355*** (0.084)	0.340*** (0.065)	0.257*** (0.061)	0.342*** (0.064)	0.383*** (0.076)
ln(1 + km to centre)	-0.035* (0.019)	0.008 (0.027)	-0.014 (0.021)	-0.049** (0.020)	-0.082*** (0.018)	-0.072*** (0.022)
ln(1 + listings within 1 km)	0.064*** (0.013)	0.100*** (0.017)	0.080*** (0.013)	0.046*** (0.012)	0.033*** (0.012)	0.037*** (0.014)
Pseudo- R^2 / R^2	0.509	0.278	0.294	0.328	0.359	0.405
Observations	2,480					
City / cell fixed effects	Yes					

Notes: Analysis sample ($N = 2,480$). Column OLS repeats M5 (HC3 standard errors); quantile columns report kernel-based standard errors. All regressions include the city / cell fixed effects and the studio and “Other” category indicators (not shown). Pseudo- R^2 is one minus the ratio of the minimised check-function loss to the loss at the unconditional quantile. ***, ** and * denote significance at the 1, 5 and 10% levels.

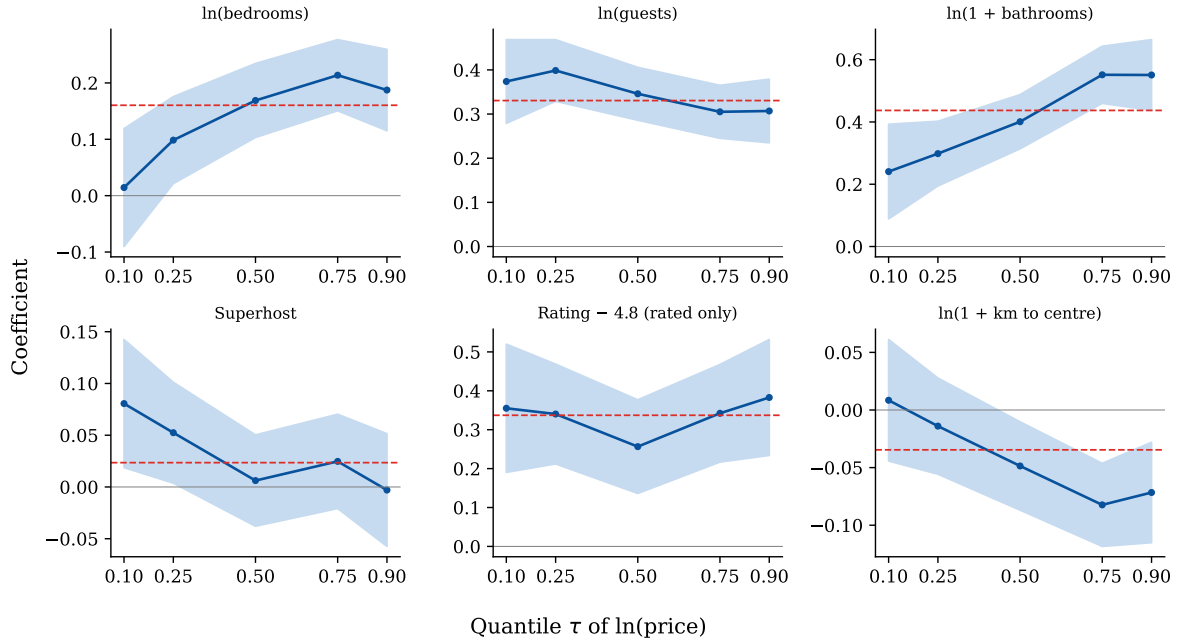


Figure 5. Quantile Regression Coefficients Across the Conditional Price Distribution

Notes: Point estimates and 95% confidence bands at $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$; the dashed red line is the OLS (M5) coefficient.

the bathroom gradient, the steepest of the three, is visible in the non-overlapping confidence bands of Figure 5. The chalet premium shows the same gradient, from 2% at the bottom to 21% at the top of the distribution.

Second, the reputation variables matter at the bottom. The Superhost badge is associated with 8% and 5% higher prices at $\tau = 0.10$ and 0.25 and with nothing above the median; the number of reviews is positively associated with price in the lower quantiles (0.042 at $\tau = 0.10$) and negatively in the upper quantiles (-0.042 at $\tau = 0.90$), the inter-quantile difference of -0.084 (standard error 0.023, $p < 0.001$) being the most significant in the table. At the lower end, reviews and the badge certify a cheap listing; at the upper end, heavily reviewed listings are high-turnover properties that are priced below the less reviewed, more exclusive ones. The rating coefficient, by contrast, is stable across quantiles (0.26 to 0.38) and its inter-quantile difference is not significant (0.03, $p = 0.86$).

Third, the location gradient strengthens with price: distance to the centre is unpriced at $\tau = 0.10$ and carries an elasticity of -0.07 to -0.08 in the upper half, while local density is priced most strongly at the bottom (0.10 at $\tau = 0.10$, 0.04 at $\tau = 0.90$). Budget listings gain from being where listings are dense, premium listings from being close to the centre.

4.6. Spatial Dependence

The residuals of the hedonic model are spatially autocorrelated. Moran's I on the M4 residuals with the five-nearest-neighbour matrix is 0.126 ($z = 10.8$) ($p < 0.001$ under 999 permutations) and it remains 0.072 ($z = 6.2$, $p < 0.001$) after the city fixed effects, so dependence exists within cities as well as between them. Table 7 reports the three spatial models estimated on the M4 regressor set. The spatial lag coefficient is $\hat{\rho} = 0.200$ (standard error 0.022) in the SAR and 0.261 (0.028) in the SDM; the spatial error coefficient is $\hat{\lambda} = 0.284$ (0.028). All three models improve substantially on OLS in log-likelihood; the SDM has the highest likelihood and the likelihood-ratio test rejects the SAR restriction $\theta = \mathbf{0}$ ($\chi^2(20) = 52.8$, $p < 0.001$), while the SEM has the lowest AIC because it adds a single parameter.

The hedonic coefficients are robust to the spatial specification: the capacity, shared-bathroom, photo and rating coefficients move by less than one standard error between OLS and any of the spatial models. The largest changes are in the variables that carry spatial information, the chalet premium (0.154 in OLS, 0.097 in the SAR) and local density (0.067 to 0.044 in the SDM), part of whose OLS coefficient is absorbed by neighbours' prices. The SDM's spatially lagged regressors are mostly insignificant; two exceptions are that a listing surrounded by high-capacity listings is priced lower ($\theta = -0.153$) and one surrounded by highly rated listings is priced lower (-0.294), conditional on its own characteristics, which is the pattern expected when neighbours are substitutes.

Table 8 translates the SAR and SDM estimates into average direct, indirect and total impacts. In the SAR the spatial multiplier is $1/(1 - \hat{\rho}) = 1.25$: the total impact of each characteristic is 25% larger than its coefficient, and a fifth of the total impact accrues to neighbouring listings through the price feedback. The direct impacts are within a few percent of the coefficients, so the OLS coefficients are good approximations of the own-listing associations. In the SDM the indirect impacts are imprecisely estimated, with two exceptions: the bathroom count has a positive spillover significant at the 5% level (total impact 0.70 against a direct impact of 0.45), and the rating has a negative spillover, significant at the 10% level, that offsets its direct impact, so the total association between the average rating of a neighbourhood and its prices is close to zero. These spillovers describe the equilibrium price surface, not a causal effect of neighbours' ratings on a listing's price.

4.7. Machine Learning Benchmarks

Table 9 compares the out-of-sample accuracy of the M5 hedonic model and five learners under random and spatially blocked cross-validation. Under random folds the M5 hedonic model reaches an out-of-sample R^2 of 0.473 and a median absolute percentage error of 28.5%; the elastic net, which regularises the same linear specification, does no better. The tree ensembles are clearly more accurate: XGBoost attains $R^2 = 0.559$, RMSE 0.450 and MdAPE 26.4%, with gradient boosting,

Table 7. Spatial Lag, Spatial Error and Spatial Durbin Models (M4 Regressors)

Variable	OLS (M4)	SAR	SEM	SDM	SDM, W_x
ln(bedrooms)	0.164*** (0.042)	0.165*** (0.034)	0.155*** (0.034)	0.151*** (0.034)	0.035 (0.068)
ln(1 + bathrooms)	0.464*** (0.067)	0.444*** (0.045)	0.437*** (0.045)	0.436*** (0.045)	0.080 (0.090)
ln(guests)	0.321*** (0.037)	0.301*** (0.031)	0.334*** (0.031)	0.333*** (0.031)	-0.153** (0.060)
Shared bathroom	-0.403*** (0.052)	-0.383*** (0.047)	-0.391*** (0.048)	-0.386*** (0.048)	-0.010 (0.087)
Chalet/cottage/cabin	0.154*** (0.033)	0.097*** (0.034)	0.136*** (0.036)	0.114*** (0.038)	-0.012 (0.058)
House	0.095*** (0.033)	0.073** (0.034)	0.096*** (0.035)	0.091*** (0.035)	-0.064 (0.065)
Condo/loft/apartment	-0.043 (0.035)	-0.041 (0.034)	-0.026 (0.034)	-0.016 (0.035)	-0.054 (0.059)
Hotel/B&B/hostel	0.122** (0.049)	0.108** (0.047)	0.102** (0.048)	0.096** (0.049)	0.078 (0.089)
ln(1 + amenities)	0.067* (0.037)	0.063* (0.033)	0.056* (0.034)	0.055 (0.034)	0.049 (0.062)
ln(1 + photos)	0.092*** (0.022)	0.080*** (0.020)	0.093*** (0.020)	0.093*** (0.020)	-0.030 (0.037)
Superhost	0.008 (0.024)	0.002 (0.023)	0.006 (0.023)	0.004 (0.023)	0.034 (0.045)
Guest Favourite	-0.034 (0.027)	-0.033 (0.026)	-0.034 (0.025)	-0.031 (0.025)	0.020 (0.052)
ln(1 + reviews)	0.015 (0.010)	0.017** (0.008)	0.015* (0.008)	0.015* (0.008)	0.005 (0.017)
Rated (star shown)	-0.033 (0.050)	-0.036 (0.043)	-0.037 (0.042)	-0.036 (0.042)	-0.012 (0.093)
Rating – 4.8 (rated only)	0.328*** (0.068)	0.315*** (0.063)	0.345*** (0.063)	0.331*** (0.063)	-0.294** (0.128)
ln(1 + km to centre)	-0.002 (0.010)	-0.006 (0.010)	-0.005 (0.013)	-0.024 (0.062)	0.021 (0.061)
ln(1 + listings within 1 km)	0.067*** (0.009)	0.053*** (0.009)	0.062*** (0.011)	0.044** (0.020)	0.008 (0.023)
ρ (spatial lag)		0.200*** (0.022)		0.261*** (0.028)	
λ (spatial error)			0.284*** (0.028)		
Log-likelihood	-1,748.7	-1,710.6	-1,702.1	-1,684.3	
AIC	3,539	3,465	3,446	3,453	
Pseudo- R^2	0.479	0.498	0.478	0.511	
Observations			2,480		

Notes: Maximum-likelihood estimates on the analysis sample with a row-standardised 5-nearest-neighbour weights matrix built on coordinates projected to kilometres. The last column reports the coefficients θ on the spatially lagged regressors of the SDM. Studio and “Other” category indicators included but not shown. Standard errors in parentheses (HC3 for OLS, asymptotic for the spatial models). ***, ** and * denote significance at the 1, 5 and 10% levels.

Table 8. Average Direct, Indirect and Total Impacts, SAR and SDM

Variable	SAR			SDM		
	Direct	Indirect	Total	Direct	Indirect	Total
ln(bedrooms)	0.166*** (0.032)	0.040*** (0.010)	0.207*** (0.040)	0.154*** (0.035)	0.098 (0.086)	0.252*** (0.095)
ln(1 + bathrooms)	0.447*** (0.042)	0.108*** (0.018)	0.555*** (0.054)	0.445*** (0.043)	0.253** (0.117)	0.698*** (0.123)
ln(guests)	0.303*** (0.030)	0.073*** (0.012)	0.376*** (0.038)	0.330*** (0.032)	-0.086 (0.075)	0.244*** (0.081)
Shared bathroom	-0.385*** (0.044)	-0.093*** (0.016)	-0.479*** (0.056)	-0.391*** (0.048)	-0.145 (0.116)	-0.536*** (0.120)
Chalet/cottage/cabin	0.098*** (0.034)	0.024*** (0.008)	0.122*** (0.042)	0.115*** (0.037)	0.024 (0.072)	0.138* (0.076)
House	0.073** (0.032)	0.018** (0.008)	0.091** (0.040)	0.089** (0.035)	-0.053 (0.088)	0.036 (0.092)
Condo/loft/apartment	-0.041 (0.035)	-0.010 (0.009)	-0.051 (0.044)	-0.019 (0.033)	-0.076 (0.081)	-0.095 (0.086)
Hotel/B&B/hostel	0.109** (0.048)	0.026** (0.013)	0.135** (0.060)	0.100** (0.047)	0.135 (0.117)	0.236* (0.124)
ln(1 + amenities)	0.064* (0.034)	0.015* (0.008)	0.079* (0.042)	0.058* (0.031)	0.082 (0.080)	0.140 (0.086)
ln(1 + photos)	0.081*** (0.021)	0.020*** (0.006)	0.100*** (0.027)	0.093*** (0.020)	-0.007 (0.047)	0.086* (0.051)
Superhost	0.002 (0.023)	0.001 (0.006)	0.003 (0.028)	0.006 (0.022)	0.046 (0.058)	0.052 (0.062)
Guest Favourite	-0.034 (0.026)	-0.008 (0.006)	-0.042 (0.032)	-0.030 (0.026)	0.016 (0.072)	-0.014 (0.079)
ln(1 + reviews)	0.017** (0.009)	0.004* (0.002)	0.021** (0.011)	0.016* (0.008)	0.012 (0.023)	0.028 (0.025)
Rated (star shown)	-0.037 (0.043)	-0.009 (0.011)	-0.046 (0.054)	-0.037 (0.042)	-0.028 (0.121)	-0.065 (0.137)
Rating – 4.8 (rated only)	0.317*** (0.058)	0.077*** (0.016)	0.394*** (0.071)	0.321*** (0.065)	-0.272* (0.165)	0.050 (0.190)
ln(1 + km to centre)	-0.006 (0.010)	-0.002 (0.003)	-0.008 (0.012)	-0.024 (0.062)	0.020 (0.061)	-0.004 (0.015)
ln(1 + listings within 1 km)	0.054*** (0.009)	0.013*** (0.002)	0.067*** (0.011)	0.045** (0.019)	0.025 (0.022)	0.070*** (0.014)

Notes: Impacts computed as in [LeSage and Pace \(2009\)](#) from the dense inverse $(\mathbf{I} - \rho\mathbf{W})^{-1}$. Standard errors (in parentheses) from 300 draws of the parameters from their estimated asymptotic distribution. ***, ** and * denote significance at the 1, 5 and 10% levels.

LightGBM and the random forest within one point of it. The gain of 0.086 in R^2 over the hedonic model is the value of the interactions and non-linearities that the semi-log specification rules out, for instance between type and capacity or between density and city.

Spatially blocked cross-validation is the sterner test, and it lowers every model by about seven points of R^2 : the hedonic model falls to 0.407, the tree ensembles to 0.48. The ranking is unchanged and the gap between the best learner and the hedonic model narrows slightly, from 0.086 to 0.076. Two conclusions follow. The random-fold figures overstate how well any of these models would price listings in an area absent from the training data, because listings in the same area share unobserved determinants of price; and the predictive advantage of the learners is not an artefact of their fitting local spatial structure more closely, since it survives when that structure cannot be exploited. A median error of 26 to 30% of the nightly price is, in either design, the practical accuracy limit of an automated valuation of Quebec short-term rentals from listing-card information alone.

Table 9. Out-of-Sample Accuracy Under Random and Spatially Blocked Cross-Validation

Model	Random 5-fold CV				Spatially blocked 5-fold CV			
	R^2	RMSE	MAE	MdAPE (%)	R^2	RMSE	MAE	MdAPE (%)
OLS (M5)	0.473	0.493	0.374	28.5	0.407	0.522	0.401	30.5
Elastic net	0.476	0.491	0.373	28.7	0.426	0.514	0.393	29.6
Random forest	0.543	0.459	0.346	26.7	0.483	0.488	0.372	28.5
Gradient boosting	0.557	0.452	0.342	26.7	0.480	0.489	0.376	29.5
XGBoost	0.559	0.450	0.339	26.4	0.483	0.488	0.375	28.8
LightGBM	0.552	0.454	0.341	26.1	0.482	0.488	0.372	28.5

Notes: Five-fold cross-validation on the analysis sample ($N = 2,480$); all models use the M5 regressors including the fixed-effect indicators. Random folds assign listings at random; blocked folds assign 20 k -means spatial clusters of listings to folds as whole blocks, so that each listing is predicted by a model that has seen no listing from its cluster. R^2 , RMSE and MAE are on the log scale; MdAPE is the median absolute percentage error on the price scale. Hyper-parameters: elastic net with cross-validated penalty; random forest 500 trees, minimum leaf 3; gradient boosting 500 trees, depth 3, learning rate 0.05; XGBoost 600 trees, depth 5, learning rate 0.03; LightGBM 600 trees, 15 leaves, learning rate 0.03.

4.8. SHAP Decomposition of the Boosted Model

Table 10 ranks the features of the XGBoost model by mean absolute SHAP value and Figure 6 shows the distribution of the contributions. Guest capacity is by far the most important feature (21% of the total attribution), followed by bedrooms (10%), local density, the aggregate of the city and cell indicators, bathrooms and the review count (9% each), the rating (7%), photos (5%), distance to the centre (5%), amenities (4%) and the shared-bathroom indicator (3%). The ranking mirrors the hedonic model in two respects: the capacity block dominates, and the badges, pets and the rated indicator are at the bottom, with the Superhost badge absent from the top sixteen. It differs in

one: the review count, whose average hedonic coefficient is zero, receives a large attribution. The beeswarm plot explains why. The SHAP values of the review count are negative for listings with many reviews and positive for listings with few, exactly the sign reversal between the lower and upper quantiles in Table 6; a linear coefficient averages the two to nothing, whereas the learner uses the variable in both directions. Together with the quantile results, this is the clearest instance in these data of an implicit price that is not a constant. SHAP values quantify how the fitted model uses each feature; they are attribution weights for associations, computed from an estimator that has no identification strategy, and they carry no causal content.

Table 10. Mean Absolute SHAP Values, XGBoost

Feature	Mean $ \phi_k $	Share of total (%)
ln(guests)	0.1865	20.9
ln(bedrooms)	0.0873	9.8
ln(1 + listings within 1 km)	0.0796	8.9
City / cell fixed effects (all)	0.0789	8.8
ln(1 + bathrooms)	0.0788	8.8
ln(1 + reviews)	0.0771	8.6
Rating – 4.8 (rated only)	0.0591	6.6
ln(1 + photos)	0.0455	5.1
ln(1 + km to centre)	0.0431	4.8
ln(1 + amenities)	0.0369	4.1
Shared bathroom	0.0270	3.0
Chalet/cottage/cabin	0.0256	2.9
Guest Favourite	0.0141	1.6
Condo/loft/apartment	0.0114	1.3
Pets allowed	0.0089	1.0
Rated (star shown)	0.0083	0.9

Notes: Exact SHAP values from the TreeExplainer for the XGBoost model fitted to the analysis sample. $|\phi_k|$ is the mean absolute contribution of feature k to the predicted log price; the share is relative to the sum over all features, including the individual fixed-effect indicators, which are reported as one aggregate row.

5. Robustness

Table 11 re-estimates the M5 specification under alternative definitions of the sample and of the fixed effects and reports the coefficients that carry the paper’s main results. The checks address three threats: that the results depend on how the stay-price regime was removed, that they are driven by a particular segment of the market, and that they depend on the spatial resolution of the fixed effects.

Table 11. Robustness of the Main Coefficients (Specification M5)

Sample / specification	Bedr.	Bathr.	Guests	Superhost	Guest Fav.	ln reviews	Rating	ln dist.	<i>N</i>	<i>R</i> ²
Baseline: nightly regime ($\hat{p}_i < 0.5$)	0.160*** (0.042)	0.437*** (0.070)	0.331*** (0.037)	0.023 (0.025)	-0.035 (0.027)	0.015 (0.010)	0.337*** (0.068)	-0.035* (0.019)	2,480	0.509
Nightly regime, strict ($\hat{p}_i < 0.25$)	0.161*** (0.042)	0.444*** (0.068)	0.329*** (0.037)	0.022 (0.025)	-0.036 (0.027)	0.017* (0.010)	0.343*** (0.067)	-0.034* (0.019)	2,475	0.516
Nightly regime, lenient ($\hat{p}_i < 0.75$)	0.160*** (0.042)	0.438*** (0.070)	0.329*** (0.037)	0.020 (0.025)	-0.039 (0.027)	0.014 (0.010)	0.354*** (0.069)	-0.031* (0.019)	2,482	0.507
Price $\leq$ \$1,000 rule	0.303*** (0.057)	0.409*** (0.083)	0.015 (0.055)	0.016 (0.030)	-0.023 (0.033)	-0.005 (0.012)	0.095 (0.101)	-0.021 (0.025)	2,541	0.282
Price $\leq$ \$600 rule	0.138*** (0.042)	0.247*** (0.072)	0.273*** (0.040)	0.020 (0.024)	-0.001 (0.027)	0.015 (0.010)	0.254*** (0.077)	-0.034* (0.020)	2,402	0.417
All priced listings (bimodal)	0.604*** (0.095)	0.281** (0.125)	-0.279*** (0.080)	-0.240*** (0.057)	-0.351*** (0.063)	-0.131*** (0.023)	0.794*** (0.176)	-0.084* (0.048)	3,456	0.158
Excluding hotels, B&Bs, hostels	0.122*** (0.040)	0.544*** (0.063)	0.327*** (0.036)	0.028 (0.026)	-0.040 (0.028)	0.015 (0.010)	0.378*** (0.074)	-0.040** (0.020)	2,335	0.513
Rated listings only	0.175*** (0.045)	0.427*** (0.075)	0.314*** (0.037)	0.004 (0.026)	-0.023 (0.027)	0.010 (0.011)	0.367*** (0.069)	-0.032 (0.021)	2,246	0.509
Montreal only	0.096 (0.086)	0.485*** (0.167)	0.313*** (0.079)	-0.053 (0.059)	0.077 (0.064)	0.112*** (0.021)	0.336** (0.147)	-0.173* (0.091)	506	0.543
Outside Montreal	0.221*** (0.048)	0.369*** (0.075)	0.322*** (0.043)	0.030 (0.027)	-0.041 (0.030)	-0.017 (0.011)	0.309*** (0.079)	-0.031 (0.019)	1,974	0.540
Trimmed 1%/99% within sample	0.181*** (0.037)	0.301*** (0.066)	0.326*** (0.032)	0.019 (0.023)	-0.012 (0.025)	0.009 (0.010)	0.342*** (0.066)	-0.031* (0.017)	2,420	0.512
Weighted by $1 - \hat{p}_i$	0.160*** (0.042)	0.442*** (0.068)	0.329*** (0.037)	0.023 (0.025)	-0.035 (0.027)	0.016 (0.010)	0.340*** (0.067)	-0.034* (0.019)	2,480	0.512
60 spatial-cell fixed effects	0.173*** (0.044)	0.459*** (0.071)	0.307*** (0.038)	0.008 (0.025)	-0.040 (0.028)	0.018* (0.010)	0.330*** (0.068)	-0.021 (0.022)	2,480	0.505

Notes: Each row re-estimates M5 (all regressors and city / cell fixed effects) on the indicated sample; HC3 standard errors in parentheses. “Bedr.”, “Bathr.” and “Guests” are the elasticities with respect to bedrooms, one plus bathrooms and guest capacity; “Rating” is the coefficient on the centred rating among rated listings; “ln dist.” is the coefficient on $\ln(1 + \text{km to centre})$. $\hat{p}_i$ is the posterior probability of the stay regime from Equation (8). The “weighted” row uses weighted least squares with weights $1 - \hat{p}_i$ on the baseline sample. The last row replaces the city / cell fixed effects by 60 *k*-means spatial cells. ***, ** and * denote significance at the 1, 5 and 10% levels.

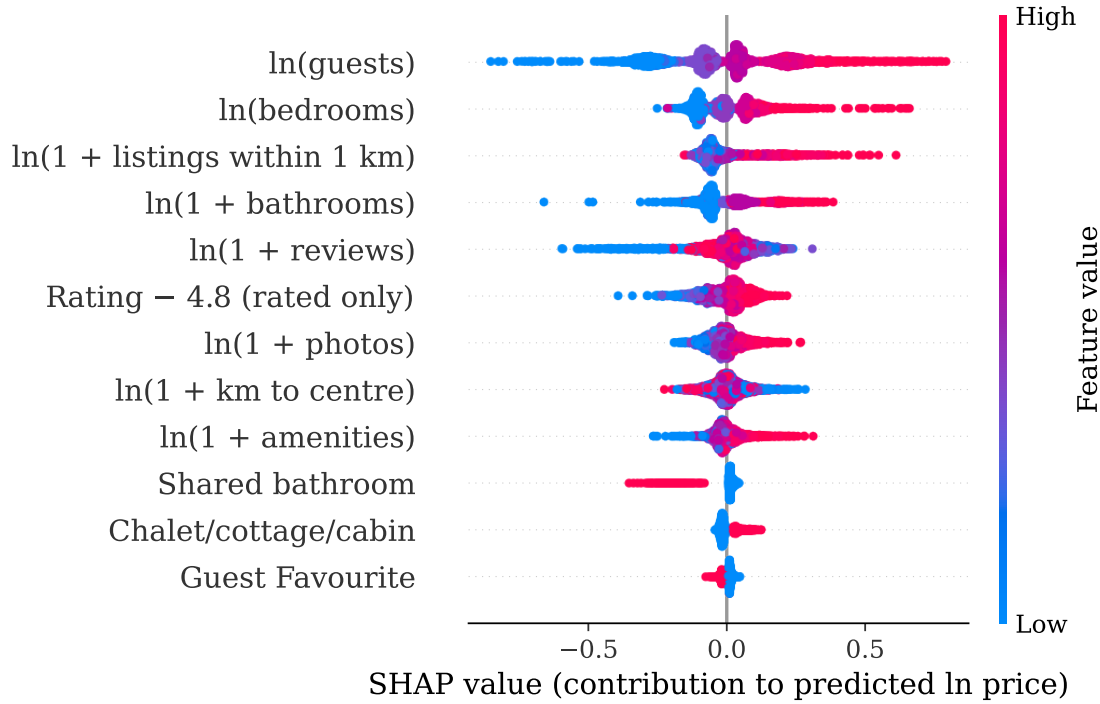


Figure 6. SHAP Summary Plot, XGBoost (Twelve Most Important Listing-Level Features)

Notes: Each point is one listing; the horizontal position is the SHAP value (contribution to predicted log price), the colour the feature value (blue low, red high). Fixed-effect indicators are excluded from the plot.

The regime classification. The baseline results are insensitive to the posterior cut-off: tightening it to 0.25 removes five listings and loosening it to 0.75 adds two, and no coefficient moves by more than 0.02. Weighting observations by $1 - \hat{p}_i$ instead of classifying them gives the same estimates. The comparison with the simple price rules is instructive. Keeping every listing priced at or below \$1,000, the rule used in the companion paper, leaves 83 stay-regime listings in the sample; because these are small, cheap-per-night units, they pull the capacity elasticity to zero (0.015 against 0.331), halve the rating coefficient and lower R^2 to 0.28. A \$600 cut-off removes all but nine of them but also discards 87 genuine nightly prices of large chalets and houses, which compresses the bathroom and rating coefficients (0.247 and 0.254) and lowers R^2 to 0.42. The full priced sample, finally, reproduces the pathology of the first version of this paper: the capacity elasticity turns negative, the badges and the review count acquire large negative coefficients, the rating coefficient doubles, and R^2 is 0.16. The model-based classification is therefore not a cosmetic choice; the price rules are biased in opposite directions and the bias is concentrated in exactly the coefficients of interest.

Composition. Excluding hotel rooms, bed-and-breakfasts and hostels, which are priced per room on a different logic, raises the bathroom elasticity to 0.54 and the rating coefficient to 0.38 and leaves the rest unchanged. Restricting the sample to rated listings, so that the rating enters for

every observation, gives the same rating coefficient (0.37) and confirms that the unrated indicator is not driving it. Splitting the sample at Montreal reveals the one substantive heterogeneity in the table. Within Montreal ($N = 506$) the bedroom elasticity is small and imprecise, the review count is positively priced (0.112, standard error 0.021) and the distance gradient is steep (-0.173 per log-kilometre); outside Montreal the bedroom elasticity is 0.22, reviews are unpriced and distance to the nearest centre does not matter. In the metropolitan market, where listings are substitutes within a dense grid, an established review record and proximity to the centre are what differentiate listings; in the resort and small-town markets, size does. The capacity, bathroom and rating coefficients are nonetheless similar in the two sub-samples, and the within-Montreal R^2 (0.54) is slightly higher than the provincial one.

Outliers and fixed effects. Trimming the top and bottom percentile of prices within the analysis sample lowers the bathroom elasticity to 0.30 and leaves the other coefficients unchanged; the largest bathroom counts belong to the most expensive chalets, so the bathroom coefficient is partly identified from the upper tail, as the quantile results already indicated. Replacing the 51 city and cell fixed effects by 60 finer spatial cells, which halves the average number of listings per group, changes no coefficient by more than 0.03 and leaves R^2 at 0.505, so the listing-level implicit prices are not sensitive to how finely location is controlled. The spatial models of Section 4.6 provide the complementary check that the coefficients are robust to modelling the dependence explicitly rather than absorbing it: with ten rather than five nearest neighbours the spatial lag coefficient is 0.215 (standard error 0.027) and the Moran statistic of the M4 residuals is 0.094 ($z = 11.4$), so the dependence is not an artefact of a particular neighbourhood definition.

In summary, the implicit prices of capacity, bathrooms, the shared bathroom, photos, the rating and local density, and the absence of a Superhost or Guest Favourite premium, hold across the regime cut-off, weighting, the exclusion of commercial lodging, trimming, finer fixed effects and alternative spatial weights. The two results that are sample-dependent are the bedroom elasticity, which is driven by the non-metropolitan market, and the review-count coefficient, which is positive in Montreal and zero elsewhere; both are consistent with the quantile patterns and are reported as heterogeneity rather than as instability.

6. Conclusion

This paper has estimated hedonic price functions for a province-wide cross-section of Airbnb listings in Quebec from a scrape of 5,000 listing cards. Its first finding concerns the data: the displayed price field mixes nightly rates with the total prices of stays of about three weeks, for roughly 28% of priced listings, concentrated among small Montreal rental units with few reviews. A two-regime

mixture of hedonic regressions separates the two populations sharply and leaves 2,480 listings with nightly prices. Every estimate in the first version of this paper was computed on the mixture; none survives, and the magnitudes reported there (a bedroom premium above 200%, an 11% premium per tenth of a rating point, negative “amenity index” effects) were artefacts of the contamination and of regressors that should not have been constructed. We have documented these corrections in full so that the lesson is usable by others who work with scraped platform prices: the bimodality was visible in the first summary table, and a ratio of medians by bedroom count is enough to diagnose it.

On the nightly-price sample the hedonic model behaves as the literature leads one to expect. Capacity, bathrooms and bedrooms, the shared-bathroom indicator that stands in for the unobserved room type, the number of photos and the displayed rating are the priced attributes; title categories matter once location is held fixed, with chalets and houses priced about 8 to 10% above rental units of the same capacity; local density is positively priced within cities. The Superhost and Guest Favourite badges carry no premium conditional on capacity, type and location, and the number of reviews is unrelated to price on average. The specification with city fixed effects explains half of the variance of log prices, with a root mean squared error of 0.475 log points, and its coefficients are stable to the choice of fixed-effect unit, to clustering, to the regime cut-off and to the exclusion of commercial lodging or of Montreal.

The three complementary estimators each add something the linear model does not show. Quantile regression reveals that cheap listings are priced per guest and certified by reviews and the Superhost badge, whereas expensive listings are priced per room and per bathroom, closer to the centre, and are cheaper when heavily reviewed; the inter-quantile differences for bedrooms, bathrooms and review counts are statistically significant. The spatial models find moderate spatial dependence ($\hat{\rho} \approx 0.20\text{--}0.26$, $\hat{\lambda} \approx 0.28$) and a spatial multiplier of about 1.25, but the direct impacts are within a few percent of the OLS coefficients, so the hedonic estimates are not materially biased by the omitted spatial lag. The machine-learning benchmarks improve the out-of-sample R^2 from 0.47 to 0.56 under random cross-validation and from 0.41 to 0.48 when whole areas are held out; the seven-point loss under spatial blocking applies to every model, and the SHAP decomposition of the best learner ranks capacity first and the badges last, as the hedonic coefficients do, and attributes to the review count a weight that the linear model cannot, because its association with price changes sign along the distribution. The flexible models refine the hedonic surface rather than contradict it, and a median pricing error of about one quarter of the nightly price is the practical limit of valuation from card-level information.

Limitations. The price is an asking price displayed on a single day and excludes the cleaning fee; the data are a cross-section, so seasonality and within-listing dynamics are unobserved and all coefficients are conditional associations rather than marginal willingness to pay. The room type, the host identifier and the amenity list were not scraped, so professionalisation and specific amenities

cannot be priced, and the stay-regime listings cannot be converted to nightly rates. The scrape reflects the platform's search results on the collection dates and is denser in urban and resort markets.

Future work. A repeated scrape that records the search dates, the stay length behind each displayed price, the room type and the host identifier would remove the main measurement problems and allow panel estimation of within-listing price responses to reviews, badges and seasons. Linking listings to realised bookings or to the calendar would replace asking prices with revenue. Finally, the Quebec market offers a natural setting to study how Montreal's registration and primary-residence rules shift the composition of listings and their prices, a question that the cross-sectional stay-price pattern documented here already hints at.

References

- Anselin, Luc**, *Spatial Econometrics: Methods and Models*, Dordrecht: Kluwer Academic Publishers, 1988.
- Bishop, Kelly C., Nicolai V. Kuminoff, H. Spencer Banzhaf, Kevin J. Boyle, Kathrine von Gravenitz, Jaren C. Pope, V. Kerry Smith, and Christopher D. Timmins**, “Best Practices for Using Hedonic Property Value Models to Measure Willingness to Pay for Environmental Quality,” *Review of Environmental Economics and Policy*, 2020, 14 (2), 260–281.
- Breiman, Leo**, “Random Forests,” *Machine Learning*, 2001, 45 (1), 5–32.
- Cai, Yuan, Yongbo Zhou, Jianyu Ma, and Noel Scott**, “Price Determinants of Airbnb Listings: Evidence from Hong Kong,” *Tourism Analysis*, 2019, 24 (2), 227–242.
- Cameron, A. Colin and Douglas L. Miller**, “A Practitioner’s Guide to Cluster-Robust Inference,” *Journal of Human Resources*, 2015, 50 (2), 317–372.
- Chen, Tianqi and Carlos Guestrin**, “XGBoost: A Scalable Tree Boosting System,” in “Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining” 2016, pp. 785–794.
- Chen, Yong and Karen Xie**, “Consumer Valuation of Airbnb Listings: A Hedonic Pricing Approach,” *International Journal of Contemporary Hospitality Management*, 2017, 29 (9), 2405–2424.
- Combs, Jennifer, Danielle Kerrigan, and David Wachsmuth**, “Short-Term Rentals in Canada: Uneven Growth, Uneven Impacts,” *Canadian Journal of Urban Research*, 2020, 29 (1), 119–135.
- Deboosere, Robbin, Danielle Jane Kerrigan, David Wachsmuth, and Ahmed El-Geneidy**, “Location, Location and Professionalization: A Multilevel Hedonic Analysis of Airbnb Listing Prices and Revenue,” *Regional Studies, Regional Science*, 2019, 6 (1), 143–156.
- Elhorst, J. Paul**, “Applied Spatial Econometrics: Raising the Bar,” *Spatial Economic Analysis*, 2010, 5 (1), 9–28.
- Ert, Eyal, Aliza Fleischer, and Nathan Magen**, “Trust and Reputation in the Sharing Economy: The Role of Personal Photos in Airbnb,” *Tourism Management*, 2016, 55, 62–73.
- Friedman, Jerome H.**, “Greedy Function Approximation: A Gradient Boosting Machine,” *Annals of Statistics*, 2001, 29 (5), 1189–1232.

- Gibbs, Chris, Daniel Guttentag, Ulrike Gretzel, Jym Morton, and Alasdair Goodwill**, “Pricing in the Sharing Economy: A Hedonic Pricing Model Applied to Airbnb Listings,” *Journal of Travel & Tourism Marketing*, 2018, 35 (1), 46–56.
- , —, —, **Lan Yao, and Jym Morton**, “Use of Dynamic Pricing Strategies by Airbnb Hosts,” *International Journal of Contemporary Hospitality Management*, 2018, 30 (1), 2–20.
- Gyódi, Kristóf and Łukasz Nawaro**, “Determinants of Airbnb Prices in European Cities: A Spatial Econometrics Approach,” *Tourism Management*, 2021, 86, 104319.
- Halvorsen, Robert and Raymond Palmquist**, “The Interpretation of Dummy Variables in Semilogarithmic Equations,” *American Economic Review*, 1980, 70 (3), 474–475.
- Kennedy, Peter E.**, “Estimation with Correctly Interpreted Dummy Variables in Semilogarithmic Equations,” *American Economic Review*, 1981, 71 (4), 801.
- Koenker, Roger and Gilbert Bassett**, “Regression Quantiles,” *Econometrica*, 1978, 46 (1), 33–50.
- **and Kevin F. Hallock**, “Quantile Regression,” *Journal of Economic Perspectives*, 2001, 15 (4), 143–156.
- Kuminoff, Nicolai V., Christopher F. Parmeter, and Jaren C. Pope**, “Which Hedonic Models Can We Trust to Recover the Marginal Willingness to Pay for Environmental Amenities?,” *Journal of Environmental Economics and Management*, 2010, 60 (3), 145–160.
- Lancaster, Kelvin J.**, “A New Approach to Consumer Theory,” *Journal of Political Economy*, 1966, 74 (2), 132–157.
- LeSage, James P. and R. Kelley Pace**, *Introduction to Spatial Econometrics*, Boca Raton, FL: CRC Press, 2009.
- Lorde, Troy, Jadon Jacob, and Quinn Weekes**, “Price-Setting Behavior in a Tourism Sharing Economy Accommodation Market: A Hedonic Price Analysis of AirBnB Hosts in the Caribbean,” *Tourism Management Perspectives*, 2019, 30, 251–261.
- Lundberg, Scott M. and Su-In Lee**, “A Unified Approach to Interpreting Model Predictions,” in “Advances in Neural Information Processing Systems 30 (NIPS 2017)” Long Beach, CA 2017, pp. 4765–4774.
- , **Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee**, “From Local Explanations to Global Understanding with Explainable AI for Trees,” *Nature Machine Intelligence*, 2020, 2 (1), 56–67.

- MacKinnon, James G. and Halbert White**, “Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties,” *Journal of Econometrics*, 1985, 29 (3), 305–325.
- Malpezzi, Stephen**, “Hedonic Pricing Models: A Selective and Applied Review,” in Tony O’Sullivan and Kenneth Gibb, eds., *Housing Economics and Public Policy*, Oxford: Blackwell Science, 2003, pp. 67–89.
- Pérez-Sánchez, V. Raul, Leticia Serrano-Estrada, Pablo Marti, and Raul-Tomas Mora-Garcia**, “The What, Where, and Why of Airbnb Price Determinants,” *Sustainability*, 2018, 10 (12), 4596.
- Roberts, David R., Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann**, “Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure,” *Ecography*, 2017, 40 (8), 913–929.
- Rosen, Sherwin**, “Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition,” *Journal of Political Economy*, 1974, 82 (1), 34–55.
- Sirmans, G. Stacy, David A. Macpherson, and Emily N. Zietz**, “The Composition of Hedonic Pricing Models,” *Journal of Real Estate Literature*, 2005, 13 (1), 1–44.
- Teubner, Timm, Florian Hawlitschek, and David Dann**, “Price Determinants on Airbnb: How Reputation Pays Off in the Sharing Economy,” *Journal of Self-Governance and Management Economics*, 2017, 5 (4), 53–80.
- Tong, Baldwin and Ulrich Gunter**, “Hedonic Pricing and the Sharing Economy: How Profile Characteristics Affect Airbnb Accommodation Prices in Barcelona, Madrid, and Seville,” *Current Issues in Tourism*, 2022, 25 (20), 3309–3328.
- Wachsmuth, David and Alexander Weisler**, “Airbnb and the Rent Gap: Gentrification Through the Sharing Economy,” *Environment and Planning A: Economy and Space*, 2018, 50 (6), 1147–1170.
- Wang, Dan and Juan L. Nicolau**, “Price Determinants of Sharing Economy Based Accommodation Rental: A Study of Listings from 33 Cities on Airbnb.com,” *International Journal of Hospitality Management*, 2017, 62, 120–131.
- Zervas, Georgios, Davide Proserpio, and John W. Byers**, “A First Look at Online Reputation on Airbnb, Where Every Stay Is Above Average,” *Marketing Letters*, 2021, 32 (1), 1–16.

Zou, Hui and Trevor Hastie, “Regularization and Variable Selection via the Elastic Net,” *Journal of the Royal Statistical Society: Series B*, 2005, 67 (2), 301–320.

A. Variable Definitions

Table [A.1](#) defines every variable used in the paper and indicates whether it comes from a structured field of the scrape or is parsed from the card title.

B. Additional Results

Table [A.2](#) reports a logistic regression of the mixture classification on listing characteristics, estimated on the full priced sample, which describes the profile of listings whose displayed price is a stay price. Table [A.3](#) reports variance inflation factors for the M4 regressors.

Table A.1. Variable Definitions

Variable	Definition	Source
<i>Dependent variable</i>		
$\ln p$	Natural log of the displayed price (CAD), winsorised at the 1st/99th percentiles of the priced sample	field
<i>Structural</i>		
$\ln$ bedrooms	Log of the bedroom count; studios recoded to one	field, title
$\ln(1 + \text{bathrooms})$	Log of one plus the bathroom count; zeros repaired from the title; half bath = 0.5	field, title
$\ln$ guests	Log of the guest capacity (minimum one)	field
Shared bathroom	1 if the title reads “shared bath”	title
Studio	1 if the title reads “Studio”	title
Title category	Six classes of the 44 host-chosen labels (reference: Rental unit)	title
<i>Presentation</i>		
$\ln(1 + \text{amenities})$	Log of one plus the number of amenities listed	field
$\ln(1 + \text{photos})$	Log of one plus the number of photos	field
Pets allowed	1 if pets are allowed	field
<i>Host and reputation</i>		
Superhost	1 if the host holds the Superhost badge	field
Guest Favourite	1 if the listing holds the Guest Favourite badge	field
$\ln(1 + \text{reviews})$	Log of one plus the number of reviews	field
Rated	1 if a star rating is displayed (0 for “New” and unrated cards)	title
Rating – 4.8	Displayed rating minus 4.8, set to zero for unrated listings	title
<i>Location</i>		
$\ln(1 + \text{km to centre})$	Log of one plus the great-circle distance (km) to the nearest of 22 reference municipalities	coordinates
$\ln(1 + \text{listings within 1 km})$	Log of one plus the number of other priced listings within 1 km	coordinates
Montreal	1 if the municipality is Montreal	field
City / cell fixed effects	Indicator for the municipality (if ≥ 15 listings) or for one of 25 k -means cells of the remaining listings	field, coordinates
<i>Regime classification</i>		
$\hat{p}_i$	Posterior probability of the stay-price regime from Equation (8)	estimated

Notes: “Field” denotes a structured field of the scrape; “title” denotes a value parsed from the listing-card title. The scraper’s price category, rating category and composite quality score are not used.

Table A.2. Which Listings Display a Stay Price? Logistic Regression of the Regime Classification

Variable	Coefficient	(SE)	Avg. marginal effect	(SE)
ln(bedrooms)	0.914***	(0.169)	0.147	(0.027)
ln(1 + bathrooms)	-0.270	(0.204)	-0.043	(0.033)
ln(guests)	-1.451***	(0.138)	-0.233	(0.020)
ln(1 + amenities)	0.016	(0.129)	0.003	(0.021)
Superhost	-0.417***	(0.093)	-0.067	(0.015)
Guest Favourite	-0.609***	(0.113)	-0.098	(0.018)
Rated (star shown)	0.113	(0.153)	0.018	(0.025)
ln(1 + reviews)	-0.314***	(0.037)	-0.050	(0.006)
Montreal	0.365***	(0.114)	0.059	(0.018)
ln(1 + km to centre)	-0.059	(0.040)	-0.009	(0.006)
Chalet/cottage/cabin	-0.419***	(0.152)	-0.067	(0.024)
House	-0.306**	(0.137)	-0.049	(0.022)
Condo/loft/apartment	-0.022	(0.135)	-0.003	(0.022)
Hotel/B&B/hostel	-0.977***	(0.209)	-0.157	(0.033)
Other/unconventional	-0.346	(0.274)	-0.056	(0.044)
Observations		3,456		
Pseudo- R^2		0.177		
Share in regime		0.282		

Notes: Dependent variable: 1 if the listing is classified in the stay regime ($\hat{p}_i \geq 0.5$). Robust (HC1) standard errors. Average marginal effects are averaged over the sample. Reference title category: Rental unit. ***, ** and * denote significance at the 1, 5 and 10% levels.

Table A.3. Variance Inflation Factors, Specification M4

Regressor	VIF
ln(bedrooms)	4.16
ln(1 + bathrooms)	1.80
ln(guests)	3.78
Shared bathroom	1.27
Studio	1.08
Chalet/cottage/cabin	2.54
House	1.98
Condo/loft/apartment	1.30
Hotel/B&B/hostel	1.33
Other/unconventional	1.24
ln(1 + amenities)	1.55
Pets allowed	1.11
ln(1 + photos)	1.62
Superhost	1.36
Guest Favourite	1.69
ln(1 + reviews)	1.87
Rated (star shown)	1.67
Rating – 4.8 (rated only)	1.53
ln(1 + km to centre)	2.16
ln(1 + listings within 1 km)	2.25

Notes: Variance inflation factors computed on the M4 design matrix (analysis sample, $N = 2,480$). Values below 5 indicate that collinearity among the capacity variables does not compromise the precision of the estimates.