Explaining Airbnb Prices in Madrid with Spatial Econometrics and Interactive Web Mapping

Virginia Di Mauro (255068)

Department of Sociology and Social Research, University of Trento

Abstract :

This study examines the determinants of Airbnb listing prices in Madrid using open data from Inside Airbnb and geospatial analysis methods. A reproducible pipeline for data cleaning and feature construction is implemented, a baseline hedonic model is estimated, and spatial dependence in residuals is assessed. Guided by exploratory spatial data analysis and diagnostic testing, spatial regression models (SAR and SEM) are then estimated to account for spatial spillovers and spatially correlated unobservables. In our sample, spatial specifications are associated with lower residual spatial autocorrelation and more stable coefficient diagnostics relative to OLS, underscoring the relevance of explicitly modeling geographic structure in short-term rental markets. Finally, an interactive web map is provided to visualize key variables and model outputs (including residual patterns), enabling transparent result inspection and scenario exploration through filters and thresholds.

Keywords: *Airbnb, Madrid, spatial econometrics, SAR model, SEM model, interactive web map.*

1 Introduction

Short-term rental platforms have become a prominent component of the accommodation market in many tourist cities, increasing interest in how listing prices are formed and how they vary across urban space. In the case of Airbnb, the listed price can be understood as the outcome of multiple factors operating simultaneously: characteristics of the accommodation itself (such as capacity and available amenities), signals related to hosts and platform mechanisms, and the broader neighborhood context in which the listing is embedded. Because these factors are not uniformly distributed across a city, substantial spatial heterogeneity in prices is typically observed, with high- and low-price areas often emerging in ways that reflect the underlying urban structure and tourism geography.

A distinctive feature of short-term rental markets is that observations are inherently geographic. Listings are points in space, situated within neighborhoods and shaped by access to attractions, services, transportation infrastructure, and other location-specific conditions. Consequently, nearby listings may share similar demand environments and may also face common competitive pressures. This geographic embedding has two important implications for empirical analysis. First, models that only rely on listing-level attributes risk overlooking a key portion of price variation that is attributable to location and neighborhood context. Second, because geographically proximate listings tend to be exposed to similar conditions, prices and model residuals may display spatial dependence, so that the standard assumption of independent observations becomes less plausible.

These considerations motivate a careful analytical approach that combines attribute-based modeling with explicit attention to spatial structure. In this report, Airbnb prices in Madrid are studied through a sequence that starts from a baseline price model and then evaluates whether spatial patterns remain in the residuals. When spatial dependence is detected, spatial regression specifications can be used to account for the geographic structure of the data and to support more coherent interpretation of model outputs. Beyond model estimation, the spatial nature of the problem also calls for representation tools that make geographic patterns visible and inspectable, helping readers assess whether the empirical results align with the spatial distribution of listings and urban context.

Building on this motivation, the project develops an empirical workflow that treats Airbnb pricing as an attribute-based problem embedded in geographic space. Listing characteristics and neighborhood context are integrated within a hedonic specification, and the extent to which spatial structure persists beyond observed covariates is evaluated through residual-based diagnostics. When spatial dependence is present, spatial regression formulations are adopted to represent either spillover dynamics among nearby listings or spatially correlated unobservables, with the aim of producing estimates that remain interpretable while being empirically coherent with the observed spatial patterns.

2 Literature review

Airbnb pricing has been examined through two tightly connected lenses: hedonic pricing theory, where nightly rates are interpreted as the implicit valuation of a bundle of observable attributes, and spatial analysis, where prices are shaped by the geographic embedding of listings in an urban system of amenities, accessibility, and localized market conditions. In the hedonic tradition, price is modeled as a function of characteristics that guests value (e.g., size/capacity, amenities, dwelling type), consistent with the foundational logic that heterogeneous goods can be decomposed into attribute bundles with implicit prices (Rosen, 1974). However, short-term rental markets are rarely "aspatial": Airbnb supply is spatially concentrated and demand drivers (tourism intensity, neighborhood attractiveness, and accessibility to points of interest) tend to vary systematically across the city, meaning that location operates not merely as a control but as a core mechanism through which willingness to pay is formed (Chica-Olmo et al., 2020).

Empirically, a recurring finding is that determinants can be organized into (i) structural/physical listing characteristics, (ii) host and platform-related signals (including professionalism proxies and status badges), (iii) reputation indicators (ratings, review volume, perceived quality), and (iv) neighborhood and locational attributes; while structural variables often show strong associa-

tions with prices, the role of reputational signals can be context-dependent and partly mediated by market maturity and local demand composition (Chen et al., 2017).

Spatial-hedonic work further emphasizes that a large share of variation is tied to the urban environment: introducing neighborhood factors and GIS-derived accessibility variables can substantially improve explanatory power relative to specifications based only on listing attributes (Chica-Olmo et al., 2020). At the same time, the relevant "location effect" is not always well summarized by a single distance-to-center proxy, especially in polycentric cities or tourism systems structured around multiple anchors (e.g., historic cores, business areas, transport hubs, major attractions); accordingly, recent studies advocate representing accessibility through multiple reference points and richer spatial covariates, including proxies of urban form and environmental context that may capture additional components of visitor experience and willingness to pay (Chica-Olmo et al., 2020).

Beyond contextual amenities, spatial dependence can also arise from market interaction: nearby listings face similar competitive pressure and can influence each other through strategic pricing responses or imitative behavior, producing local co-movement in prices that is consistent with "peer-contagion" or spillover mechanisms (Chen et al., 2017; Chica-Olmo et al., 2020).

This local structure has two methodological implications that are central to empirical modeling. First, exploratory spatial data analysis (ESDA), including global measures such as Moran's I, is used to detect clustering and to test whether OLS residuals retain geographic structure; when residual spatial autocorrelation persists, OLS inference may become unreliable because the independence assumption is violated (Bernardi et al., 2023). Second, spatial econometric specifications provide principled ways to represent dependence: SAR models capture outcome spillovers through a spatial lag of the dependent variable (consistent with endogenous interaction/competition effects), while SEM models capture spatially structured unobservables through the error process (consistent with omitted neighborhood fundamentals), and both approaches have been applied in spatial-hedonic Airbnb studies to improve diagnostics and interpretability (Chica-Olmo et al., 2020; Bernardi et al., 2023).

Related methodological perspectives, including multilevel approaches, similarly stress that neighborhood-level structure can alter estimated relationships if it is not adequately represented, reinforcing the need to account for hierarchical or spatially patterned heterogeneity (Deboosere et al., 2019). Finally, spatial modeling changes how results are interpreted: covariates can generate direct effects on a listing and indirect effects that propagate through the spatial structure, which is especially relevant when the goal is to understand local externalities and market interaction rather than only predictive performance (Bernardi et al., 2023).

Taken together, prior work motivates a hedonic-to-spatial workflow for city-specific analysis, where ESDA-driven diagnostics justify the move from OLS to SAR/SEM and where results can be meaningfully interrogated in geographic space, acknowledging that the strength and form of location effects remain context-dependent across cities (Chen et al., 2017; Chica-Olmo et al., 2020; Deboosere et al., 2019; Bernardi et al., 2023).

2.1 Research question

Building on this literature, the project focuses on Madrid as a case where listing attributes and neighborhood context jointly shape prices, and where spatial dependence may reflect both shared local fundamentals and competitive interaction among nearby hosts. The empirical objective is therefore not only to estimate attribute and location associations, but also to test whether explicitly accounting for spatial dependence improves model adequacy as reflected by residual spatial structure and coefficient diagnostics.

The study addresses the following question:

> *RQ: To what extent are Airbnb listing prices in Madrid associated with listing characteristics and location, and does explicitly modeling spatial dependence (via SAR/SEM models) improve residual spatial behavior and coefficient diagnostics relative to a baseline OLS hedonic model?*

3 Data and Study Area

This study focuses on Madrid and uses an Inside Airbnb snapshot to define both the study boundary and the primary observational units. Rather than relying on an official administrative boundary dataset, the study area is operationalized through the neighborhood polygons distributed with the snapshot (`neighbourhoods.geojson`). These polygons provide a consistent within-city partition for aggregation and spatial diagnostics, and they align with the spatial support of the listings released in the same snapshot. In the processing workflow, the neighborhood layer used for analysis preserves the original polygon geometries and adds derived aggregate attributes. The number of polygons is 128.

Madrid is one of the largest short-term rental markets in Europe and exhibits a strong center–periphery structure: tourist demand and amenities concentrate around the historic core and major activity hubs, while peripheral districts tend to host a lower-density supply. This urban structure motivates an explicitly spatial perspective, since listings are not uniformly distributed and nearby areas often share both amenities and market conditions.

The overall spatial extent of the boundary layer is summarized by the bounding box:

$$\begin{aligned}\text{bbox}_{\text{EPSG:4326}} &= (\min x, \min y, \max x, \max y)\\ &= (-3.888963,\ 40.312065,\ -3.518051,\ 40.643271).\end{aligned}$$

This bounding box is computed from `total_bounds` of the boundary layer. The same extent is observed in both the original and processed neighborhood files, suggesting that preprocessing did not alter the overall study-area extent.

3.1 Data sources, licenses, and ethics

The main data source is Inside Airbnb (Madrid snapshot, 14 September 2025). The repository uses both the detailed and summary releases made available for this snapshot, specifically: (i) detailed listings, (ii) detailed calendar, (iii) detailed reviews, (iv) listings summary, (v) reviews summary, (vi) neighborhoods table, and (vii) neighborhoods GeoJSON. Reproducible download and unzip commands (curl + gunzip), together with local naming conventions that avoid collisions between detailed and summary files,

are documented in the project documentation (README and reproducibility document; see also the reproducibility appendix for execution details).

From an ethics and data-use perspective, Inside Airbnb redistributes platform-derived information originating from publicly accessible listing pages. In this project, the data are used exclusively for scientific and educational purposes. No attempt is made to re-identify individuals, and results are communicated through aggregate summaries and model-based diagnostics (tables, residual checks, and maps), rather than through any listing-level disclosure intended to profile hosts or guests. Interpretation is explicitly conditioned on well-known limitations of platform-derived datasets, including selection effects (non-random participation across space), host behavioral endogeneity (supply and pricing decisions co-evolve with neighborhood conditions), and potential coverage or reporting biases.

No dedicated external auxiliary geospatial datasets (e.g., OSM POI layers or metro shapefiles) are versioned in this repository. Spatial structure is therefore captured through listing coordinates, neighborhood polygons, and a distance-to-CBD proxy (Puerta del Sol) derived within the workflow. When used in modeling, distance is computed after reprojection to a metric CRS (EPSG:25830), hence expressed in meters.

3.2 Preprocessing and feature engineering

Data preparation follows a deterministic sequence that separates basic data hygiene from modeling-oriented transformations:

- **Basic normalization and cleaning.** Column names are normalized and price strings are parsed using a currency-aware routine to handle heterogeneous formats. Observations with missing coordinates are removed to ensure spatial completeness. Finally, full-row duplicates and duplicate listing IDs are filtered to avoid repeated units in estimation and mapping.
- **Price processing (deterministic order).** Because raw nightly rates are strongly right-skewed and sensitive to extreme values, the workflow applies a two-stage mitigation before transformation. First, hard plausibility thresholds remove prices < 10 and $> 10{,}000$. Second, winsorization is applied at the 0.5% and 99.5% quantiles to reduce leverage from remaining extremes. Third, log(price) is computed from the winsorized price and used as the main modeling target.
- **Target and covariates.** The target is log(price). Covariates are constructed to represent structural capacity, host attributes, and reputation signals. Room type is dummy-encoded, and neighborhood effects are included where required by specification.
- **Missing-data handling.** Median imputation is used for bedrooms, beds, and bathrooms to avoid discarding otherwise valid listings due to sparse missingness in capacity fields. Boolean mapping is applied to `host_is_superhost` and `instant_bookable`. The final estimation samples are then obtained through complete-case filtering over the target and the selected model covariates.

Unit of analysis. The workflow uses multiple spatial supports, each aligned with a specific analytical purpose:

- **Listing points** are the primary units for OLS/SAR/SEM estimation and residual diagnostics, capturing fine-grained within-city variation.
- **Neighborhood polygons** support areal aggregation (e.g., neighborhood-level summaries) and neighborhood-based spatial diagnostics.
- **Grid cells** are used for webmap aggregation to reduce overplotting and provide a readable exploratory layer. In the current webmap implementation, the grid is defined in EPSG:4326 with `grid_size = 0.05` degrees; it is used for visualization/aggregation in the web map and not for metric computations.

Sample-flow audit trail. To make selection effects transparent, the pipeline saves an audit table tracking how the analytic sample evolves. The reported transitions are: initial $N = 24{,}987$; after valid price parsing $N = 18{,}940$; after winsorization (0.5%–99.5%) $N = 18{,}765$; and after complete-case feature filtering $N = 15{,}641$. Accordingly, two analytical sample definitions coexist across pipeline components:

$$N = 18{,}940 \quad \text{(price-valid sample)},$$
$$N = 15{,}641 \quad \text{(model-complete sample)}.$$

This distinction is retained explicitly throughout the analysis so that diagnostic results can be interpreted with respect to the underlying inclusion criteria.

3.3 EDA and quality checks

Exploratory analysis and quality assurance are designed to ensure that spatial patterns are interpretable and that subsequent modeling rests on coherent geometric and statistical foundations. In brief, the workflow (i) enforces an explicit CRS policy, using EPSG:4326 for web output and visualization and EPSG:25830 for metric operations; (ii) verifies geometry validity for the neighborhood polygons; (iii) checks join coverage for the point-in-polygon assignment of listings to neighborhoods; and (iv) documents missingness, duplicates, and outlier handling through stepwise counts.

Table 1 documents the sample-flow audit trail and a full-versus-subset consistency check for global spatial autocorrelation. Figures 1–2 summarize the most relevant exploratory patterns: Figure 1 highlights the strong right-skew that motivates transformation and outlier mitigation, while Figure 2 provides a first view of spatial heterogeneity in price levels across the city. Diagnostic comparisons across model families are presented later alongside the main results (Section **??**).

4 Methods

4.1 Hedonic baseline (OLS)

We model Airbnb nightly prices with a semi-log hedonic specification estimated on the centralized modelling sample (`data/processed/model_sample.parquet`). The dependent variable is

$$y_i = \ln(\text{price\_winsorized}_i),$$

where `price_winsorized` denotes the nightly price after deterministic winsorization at the 0.5% and 99.5% quantiles (performed prior to the log transform).

Table 1
Sample-flow audit trail and full-versus-subset consistency check. The pipeline yields a price-valid sample (N = 18,940) and a model-complete sample (N = 15,641); global Moran's I remains stable across these definitions, suggesting that observed spatial clustering is not an artifact of the final complete-case restriction.

Block	Metric	Full / stage	Subset
Sample flow	Initial load	24,987	—
Sample flow	After valid price parsing	18,940	—
Sample flow	After winsorization (0.5%–99.5%)	18,765	—
Sample flow	After complete-case feature filtering	—	15,641
Consistency	Moran's I (listing, kNN-8)	0.0925	0.092151
Consistency	Moran's p-value (listing, kNN-8)	< 0.001	0.001

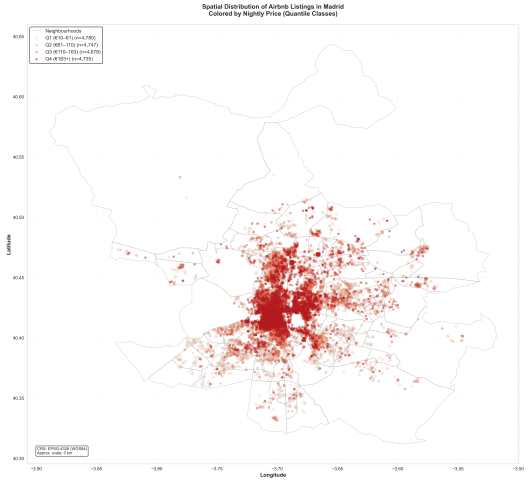

Figure 1. Distribution of listing prices in the Madrid snapshot. The pronounced right tail motivates log-transformation and outlier mitigation prior to model estimation.

The baseline parsimonious equation is:

$$\ln(\text{price\_winsorized}_i) = \alpha + X_i\beta + \varepsilon_i. \tag{1}$$

In the final parsimonious specification used for spatial models (K = 14, excluding the intercept), X_i includes exactly:

`{accommodates, bedrooms, beds, bathrooms, minimum_nights, host_is_superhost, host_listings_count, number_of_reviews, review_scores_rating, instant_bookable, dist_cbd_km, room_type_Hotel room, room_type_Private room, room_type_Shared room}`

Room-type coefficients are interpreted relative to the omitted reference category (the base room type in the one-hot encoding, i.e., the category not listed among the dummies).

A second OLS robustness model includes neighbourhood fixed effects via `neigh_*` dummies (K = 141 in saved outputs).

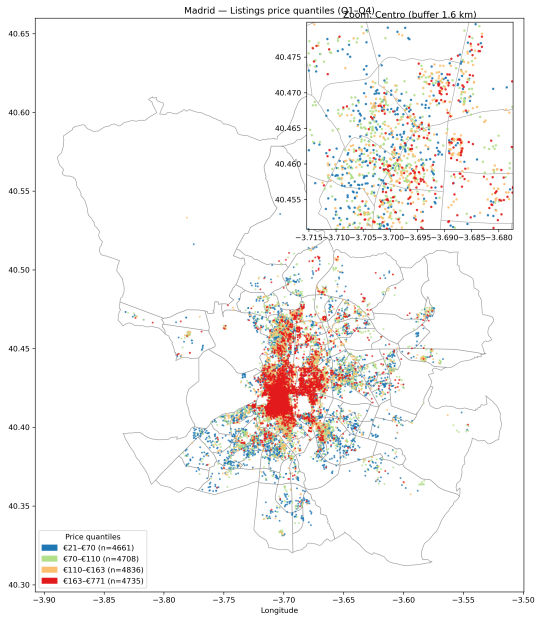

Figure 2. Spatial distribution of listing-level prices in Madrid. Higher-price clusters concentrate in central areas, while peripheral neighborhoods show lower typical prices and a sparser supply footprint.

OLS inference is computed with heteroskedasticity-robust standard errors (`cov_type='HC1'`). The setup is cross-sectional and associational, not causal.

4.2 Spatial dependence assessment (ESDA)

ESDA is implemented as *global* Moran's I on model residuals (rather than on ln(price)) in the current reproducible outputs, because residual-based diagnostics directly assess the OLS independence assumption. For a centered variable z and weights matrix $W = [w_{ij}]$:

$$I = \frac{N}{S_0}\frac{\sum_i \sum_j w_{ij} z_i z_j}{\sum_i z_i^2}, \qquad S_0 = \sum_i \sum_j w_{ij}. \tag{2}$$

Three residual-based Moran checks are produced:

1. neighbourhood-level Moran on aggregated OLS residual means with Queen contiguity (`morans_results.csv`);
2. listing-level Moran consistency check on OLS residuals for the subset workflow (`morans_results_subset.csv`);
3. post-fit Moran comparison across OLS/SAR/SEM residuals (`morans_postfit.csv`, `morans_postfit_compare.png`).

Local Moran/LISA statistics are not part of the core reported workflow in this section.

4.3 Spatial weights matrix W

The project uses two implemented weight constructions, aligned with the unit of analysis:

- **Listing-level models/diagnostics:** k-nearest neighbours (k = 8);

- **Neighbourhood-level ESDA:** Queen contiguity on neighbourhood polygons (a standard choice for administrative areal units).

For listing-level kNN, points are reprojected from EPSG:4326 to EPSG:25830 before building W, then row-standardized (`w.transform='r'`):

$$w_{ij}^{*} = \frac{w_{ij}}{\sum_j w_{ij}}. \tag{3}$$

Saved runs report zero islands for $k = 8$.

No systematic sensitivity analysis across multiple k values or distance thresholds is currently stored in reproducible outputs. Therefore, the reported spatial results are conditional on the $k = 8$ specification.

CRS policy. EPSG:4326 is used for web/visual outputs and GeoJSON interoperability. EPSG:25830 is used for metric-distance operations and listing-level spatial weights. This policy is explicitly implemented in the spatial scripts.

4.4 Spatial econometric models

Residual diagnostics indicate remaining spatial dependence after OLS, so the project estimates SAR and SEM models using GMM estimators from `spreg` on the parsimonious covariate set ($K = 14$).

SAR (spatial lag, GMM).

$$\mathbf{y} = \rho W\mathbf{y} + X\beta + \varepsilon. \tag{4}$$

ρ is estimated and reported (with p-value) in model and coefficient tables.

SEM (spatial error, GMM).

$$\mathbf{y} = X\beta + \mathbf{u}, \qquad \mathbf{u} = \lambda W\mathbf{u} + \xi. \tag{5}$$

λ is estimated and reported analogously.

Operational diagnostics and spatial specification. To make the transition from the hedonic baseline to spatial econometric models explicit, we report LM diagnostics in Table 2. The joint significance of LM-lag, LM-error, and their robust counterparts supports the presence of spatial dependence in OLS residuals and motivates estimating both SAR and SEM. For reproducibility, Table 3 summarizes the two weights structures actually implemented in the pipeline (listing-level kNN and neighbourhood-level Queen), including CRS and row-standardization choices. Finally, Figure 3 provides a visual diagnostic of neighbourhood-level residual autocorrelation under Queen contiguity, complementing the numeric Moran's I results and strengthening the ESDA-to-modeling narrative.

LM diagnostics and model comparison. LM-lag, LM-error and robust variants are computed on OLS residuals and saved in `lm_tests_modelB_listing_knn8.csv`. Main model comparison outputs include:

- R^2/pseudo-R^2,
- spatial parameter (ρ or λ) with p-value,
- σ^2,

Table 2
LM diagnostics on OLS residuals (listing level, kNN-8).

Test statistic	Value	p-value
LM-lag	1007.049	< 0.001
Robust LM-lag	506.987	< 0.001
LM-error	688.152	< 0.001
Robust LM-error	188.091	< 0.001

Table 3
Implemented spatial weights specifications used in diagnostics and models.

Analysis level	Weights type	CRS	Row-std	Notes
Listing (OLS/SAR/SEM)	kNN ($k = 8$)	EPSG:25830	Yes	Main modelling setup
Neighbourhood (ESDA)	Queen contiguity	EPSG:4326	Yes	Residual means by neighbourhood

- post-fit Moran's I on residuals.

AIC/log-likelihood are available for OLS; for SAR/SEM GMM these likelihood-based criteria are not reported (left empty in saved comparison tables). Therefore, model preference is discussed primarily in terms of residual spatial autocorrelation reduction and diagnostic evidence rather than likelihood-based fit. Direct/indirect/total impact decomposition is not reported in the current pipeline.

5 Summary of results

5.1 OLS baseline

On the main modelling sample ($N = 18{,}940$), adding location controls in OLS improves fit from Model A to Model B: R^2 increases from 0.5047 to 0.5735, while AIC decreases from 27309.95 to 24734.11 (Table 4). This comparison indicates that location variables capture a substantial part of price heterogeneity that is not explained by structural and host characteristics alone. In practical terms, moving from Model A to Model B improves explanatory power in a way that is consistent with the role of spatial context in the Madrid market.

5.2 Spatial diagnostics

LM diagnostics on the listing-level OLS specification with kNN ($k = 8$) reject the null of no spatial dependence for both lag and error forms: LM-lag = 1007.05 ($p < 0.001$), robust LM-lag = 506.99 ($p < 0.001$), LM-error = 688.15 ($p < 0.001$), robust LM-error = 188.09 ($p < 0.001$) (Table 5). These results imply that residual dependence across nearby listings is statistically meaningful and should not be ignored in model specification. LM tests are reported for Model B; SAR/SEM results below refer to the parsimonious structural specification.

5.3 Spatial models and residual autocorrelation

For the same modelling sample ($N = 18{,}940$) and weights ($k = 8$), SAR and SEM are estimated on the parsimonious structural covariate set. SAR estimates $\rho = 0.2022$ ($p < 0.001$) and reduces residual Moran's I to 0.0715. SEM estimates $\lambda = 0.4234$ ($p < 0.001$), but residual Moran's I is 0.1716. By comparison, OLS residual Moran's I is 0.1647 (Tables 6 and 7). The key interpretation is that, in this pipeline, both spatial parameters are significant, but only SAR shows a marked reduction in residual spatial autocorrelation relative to OLS. SEM captures spatial error structure through λ, yet it does not reduce residual Moran's I relative to OLS.

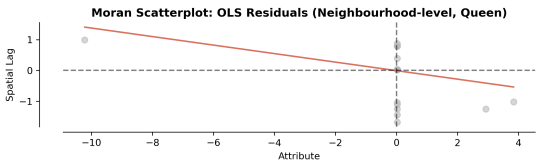

Figure 3. Moran scatterplot for neighbourhood-level mean OLS residuals with Queen contiguity weights.

Table 4
OLS comparison (Model A vs Model B, N = 18,940).

Metric	Model A	Model B
N	18940	18940
R^2	0.5047	0.5735
Adj. R^2	0.5043	0.5703
AIC	27309.95	24734.11
BIC	27419.84	25848.67
Residual Std Err	0.4974	0.4631

Table 5
LM diagnostics (listing level, kNN with k = 8, N = 18,940).

Test statistic	Value
LM-lag	1007.0486
p-value LM-lag	< 0.001
Robust LM-lag	506.9873
p-value Robust LM-lag	< 0.001
LM-error	688.1523
p-value LM-error	< 0.001
Robust LM-error	188.0911
p-value Robust LM-error	< 0.001

Table 6
Main model comparison (same sample, N = 18,940, k = 8).

Model	Fit metric (R^2 or pseudo-R^2)	Spatial coeff.	Spatial coeff. p-value	Residual Moran's I
OLS parsimonious	0.5358	–	–	0.1647
SAR (GMM)	0.5651	$\rho = 0.2022$	6.22×10^{-54}	0.0715
SEM (GMM)	0.5351	$\lambda = 0.4234$	5.65×10^{-184}	0.1716

5.4 Main takeaway

Location and neighbourhood context explain a substantial share of price variation, and spatial dependence is empirically relevant. Taken together, the diagnostics and post-fit residual evidence support the use of spatial modelling rather than relying on non-spatial OLS alone. Within the tested specifications, SAR performs best in terms of residual spatial autocorrelation reduction.

6 Interactive web map

6.1 Design goals

The interactive map is implemented as an analytical diagnostic tool to inspect where model errors concentrate in space and how those patterns change across model choices. In line with the project research question, it supports three main checks:

- Where are prices systematically over- or under-predicted under OLS versus SAR?
- How do residual patterns change when switching between OLS, SAR, and the OLS–SAR difference view?
- Which filtered subsets of listings (price, room type, accommodates) are associated with high absolute residuals?

The interface exposes the variables used for these checks directly in filters and popups: `price_numeric`, `log_price`, `accommodates`, one-hot `room_type_*`, `residual_OLS`, `residual_SAR`, plus grid aggregates (`residual_OLS_mean`, `residual_SAR_mean`, `price_mean`, `count`).

6.2 Layers and interactions

The app contains two effective map layers: (i) a listing-level point layer and (ii) a grid-polygon aggregate layer.

No explicit cluster/hotspot layer is implemented in the current app version, and no dedicated neighbourhood-boundary overlay is rendered in the map view. At least one covariate dimension is still analytically available through controls and popup fields (e.g., `accommodates`, `room_type_*`, `price_numeric`).

Legend interpretation is encoded with a discrete diverging rule centered at zero residual (blue = negative residuals, gray = near zero, red = positive residuals), using fixed thresholds implemented in code. Popups expose the key fields needed for local interpretation (residual value/sign, price level, and listing attributes).

This directly supports Section 5 validation: the map allows spatial inspection of where residual concentration weakens under SAR relative to OLS, consistent with the residual-autocorrelation evidence reported in model tables.

6.3 How to run

Run from the project root:

- `bash webmap/run.sh`
- or `streamlit run webmap/app.py`

The launcher runs QA checks first; if derived files are missing, it regenerates them deterministically via `scripts/07b_extract_residuals.py` and `scripts/08_prepare_map_layers.py`.

Required inputs read by the app are:

- `data/processed/model_sample.parquet`
- `outputs/tables/residuals_for_map.csv`
- `data/processed/map_points_sample.geojson`
- `data/processed/map_grid_cells.geojson`

The app itself does not write output files during interactive use; file generation occurs only in the preprocessing scripts above.

Environment reproducibility is pinned in `environment/environment.yml` (notably `python=3.12`, `streamlit>=1.28.0`, `streamlit-folium>=0.11.0`, `folium>=0.14.0`), with deployment runtime note in `webmap/runtime.txt`.

CRS handling is explicit: visualization layers are read/displayed in EPSG:4326, while distance-based spatial weights used to pro-

Table 7
Post-fit residual Moran's I.

Model	Moran's I	p-value	z-score
OLS parsimonious	0.1647	0.001	46.70
SAR (GMM)	0.0715	0.001	19.73
SEM (GMM)	0.1716	0.001	49.81

Table 8
Layer inventory for the interactive web map.

Layer name	Geometry	Variables visualized	Source file path	CRS (viz)	Join key	Generation step
Listings sample	Points	residual_OLS / residual_SAR / OLS-SAR; popup: price, accommodates, review_scores_rating	data/processed/map_points_sample.geojson + outputs/tables/residuals_for_map.csv	EPSG:4326	listing_id	scripts/08_prepare_map_layers.py
Grid aggregates	Polygons	residual_OLS, mean |residual_SAR|, mean |ols_diff|; popup: price_mean, count	data/processed/map_grid_cells.geojson	EPSG:4326	grid_id	scripts/08_prepare_map_layers.py

duce mapped residuals are built in metric EPSG:25830 before web visualization.

7 Conclusions

This study examined Airbnb pricing in Madrid with a central question: how much of observed price variation can be explained by listing characteristics alone, and how much requires explicit treatment of spatial structure and spatial dependence. In other words, the analysis asked whether geography is only background context or a core part of the pricing patterns visible in the data.

The results point to a consistent narrative. When location controls are added, OLS fit improves from Model A to Model B. This matters because it shows that where a listing is located is not a marginal detail: neighbourhood context and accessibility information help explain price variation beyond structural listing attributes. At the same time, the OLS residuals still display spatial structure, and LM diagnostics confirm that this remaining dependence is statistically relevant. The practical implication is that part of the model error is not random in space, so a purely non-spatial specification is likely to leave geographically clustered misfit. Within the spatial models estimated in this pipeline, SAR achieves the strongest reduction in residual spatial autocorrelation relative to both OLS and SEM. In practical terms, this indicates a better alignment between model structure and the spatial organization of errors in the data.

From a substantive perspective, these findings are coherent with the behavior of a dense urban market, where nearby listings tend to share local conditions and market signals. The evidence is consistent with the importance of local context and spatial spillovers in shaping observed pricing outcomes, while remaining descriptive rather than causal. More broadly, the results suggest that incorporating spatial dependence improves the interpretation of pricing patterns when residuals are geographically structured.

The study also has clear limitations. Results depend on how spatial relations are encoded, including choices such as the W specification and kNN design. Because the data are observational, the analysis cannot remove all confounding risks, and some relevant determinants of price variation may remain unobserved.

The interactive web map supports these conclusions by showing where residuals concentrate and how those spatial patterns change across model views.

Overall, the project contributes a coherent framework that combines econometric modelling with spatial diagnostics to study Airbnb price variation in Madrid. Its main added value is methodological clarity with reproducible execution, so that the full analytical pipeline can be rerun and audited end to end.

Table 9
Implemented controls and analytical function.

Control	Data field(s)	Function	Evaluation mode
Price range slider	price_numeric	Subset listings by nightly price interval	Server-side (Streamlit rerun)
Room-type multiselect	room_type_*	Keep selected room categories	Server-side
Accommodates slider	accommodates	Subset listings by guest capacity	Server-side
Model radio	residual_OLS / residual_SAR / OLS-SAR	Toggle residual surface shown	Server-side
Residual threshold slider	|residual|	Highlight large residual magnitudes	Server-side styling
Layer toggles	points / grid	Show-hide listing and aggregate layers	Server-side render logic

References

Bernardi, Mauro and Mariangela Guidolin (2023). "The determinants of Airbnb prices in New York City: a spatial quantile regression approach". In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 72.1, pp. 104–143. DOI: 10.1093/jrsssc/qlad001.

Chen, Yuchen and Karen L. Xie (2017). "Consumer valuation of Airbnb listings: a hedonic pricing approach". In: *International Journal of Contemporary Hospitality Management* 29.9, pp. 2405–2424. DOI: 10.1108/IJCHM-10-2016-0606.

Chica-Olmo, Jorge, Juan Gabriel González-Morales, and José Luis Zafra-Gómez (2020). "Effects of location on Airbnb apartment pricing in Málaga". In: *Tourism Management* 77, p. 103981. DOI: 10.1016/j.tourman.2019.103981.

Deboosere, Robbin et al. (2019). "Location, location and professionalization: a multilevel hedonic analysis of Airbnb listing prices and revenue". In: *Regional Studies, Regional Science* 6.1, pp. 143–156. DOI: 10.1080/21681376.2019.1592699.

Rosen, Sherwin (1974). "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition". In: *Journal of Political Economy* 82.1, pp. 34–55. DOI: 10.1086/260169.

8 Additional Resources

For transparency and reproducibility, the full project repository, including code, configuration files, and instructions to reproduce the analyses and the interactive web map—is publicly available on GitHub at `virginiadimauro/geospatial-project`[1]. The repository provides step-by-step commands to (i) download and preprocess the data, (ii) run the statistical and spatial models, and (iii) generate the outputs reported in this study.

[1] https://github.com/virginiadimauro/geospatial-project

9 Reproducibility Appendix

9.1 Computational environment

All analyses were executed in Python using a pinned Conda/Micromamba environment specification.

- **Operating system (tested):** macOS (should also work on Linux; Windows recommended via WSL2).
- **Python environment:** Conda/Micromamba environment `geo`.
- **Dependency management:** `environment/environment.yml`.

To recreate the environment:

```
micromamba env create -f environment/environment.yml
micromamba activate geo
```

Version logging (recommended). To record the exact runtime versions used:

```
python --version
python -c "import sys,platform; print(platform.platform()); print(sys.version)"
python -c "import geopandas,pyproj,shapely; \
print('geopandas',geopandas.__version__); \
print('pyproj',pyproj.__version__); \
print('shapely',shapely.__version__)"
```

9.2 Data acquisition

Raw data are sourced from Inside Airbnb (Madrid snapshot, 14 September 2025). The workflow expects source files under `data/original/` and writes cleaned/processed artifacts under `data/processed/`.

Download (raw datasets + spatial boundaries). Run from project root:

```
mkdir -p data/original
cd data/original

BASE="https://data.insideairbnb.com/spain/comunidad-de-madrid/madrid/2025-09-14"

# Compressed raw tables
curl -L -O "$BASE/data/listings.csv.gz"
curl -L -O "$BASE/data/calendar.csv.gz"
curl -L -O "$BASE/data/reviews.csv.gz"

# Summary tables (correct endpoints)
curl -L -o listings_summary.csv "$BASE/visualisations/listings.csv"
curl -L -o reviews_summary.csv  "$BASE/visualisations/reviews.csv"

# Neighbourhood boundaries (tabular + geojson)
curl -L -O "$BASE/visualisations/neighbourhoods.csv"
curl -L -O "$BASE/visualisations/neighbourhoods.geojson"

cd ../..
```

Unzip.

```
gunzip -f data/original/listings.csv.gz \
         data/original/calendar.csv.gz \
         data/original/reviews.csv.gz
```

Notes on versioning. Raw files may not be fully tracked in Git due to size constraints. Reproducibility is anchored to the snapshot date (2025-09-14), the documented URLs, and the expected filenames under `data/original/`. If the upstream provider changes availability, the project can be reproduced by re-downloading the same snapshot from the same path (or by using an archived copy, if provided).

9.3 Pipeline execution

The workflow is organized in phases. Commands should be run from project root.

Phase A (authoritative): data preparation and canonical outputs. The notebook produces the processed datasets used by all downstream analyses.

```
jupyter execute notebooks/05_final_pipeline.ipynb --inplace
```

Phase B: analysis scripts (consume processed data).

```
python scripts/01_verify_spatial_data.py
python scripts/03_ols_price_analysis.py
python scripts/05_lm_diagnostic_tests.py
python scripts/04_spatial_autocorr_morans_i.py
python scripts/06_morans_i_subset_consistency_check.py
python scripts/07_spatial_models_sar_sem.py
python scripts/02_make_static_map_overview_inset.py
```

Note on order. LM diagnostics (05) depends on the OLS model outputs (03). Moran's I (04/06) is interpreted alongside the OLS/LM results, while spatial models (07) are estimated after diagnostics.

Phase C: webmap support layers (only if running the interactive map).

```
python scripts/07b_extract_residuals.py
python scripts/08_prepare_map_layers.py
```

Table 10
Pipeline steps and main input/output artifacts.

Step	Main inputs	Main outputs
05_final_pipeline.ipynb	data/original/*	data/processed/*.parquet, geojson layers
01 Verify spatial data	processed neighbourhoods/listings	QA checks (CRS, geometry validity)
03 OLS baseline	model sample	OLS tables, residual diagnostic figures
05 LM diagnostics	OLS residuals + kNN weights	LM test tables (preferred weights spec)
04 Moran's I (ESDA)	OLS residuals + spatial support	morans_results.csv, Moran figures
06 Subset consistency	model-complete subset	morans_results_subset.csv
07 SAR/SEM models	model sample + kNN weights	coeff. tables, fit comparison, post-fit Moran
07b Residual extraction	model sample + SAR/SEM fit	residuals_for_map.csv
08 Map layers	model sample + residuals	map_points_sample.geojson, map_grid_cells.geojson

Inputs and outputs (summary).

9.4 Web map: how to run

The interactive web map is implemented with Streamlit under `webmap/`.

Recommended run command.

```
bash webmap/run.sh
```

Direct run alternative.

```
micromamba run -n geo streamlit run webmap/app.py
```

Expected artifacts. The web map reads:

- `data/processed/model_sample.parquet`,
- `outputs/tables/residuals_for_map.csv`,
- `data/processed/map_points_sample.geojson`,
- `data/processed/map_grid_cells.geojson`.

9.5 Reproducibility checklist

- **CRS policy:** EPSG:4326 for web outputs/visualization; EPSG:25830 for metric operations (distance/area/weights). Distances/areas are never computed in geographic degrees.
- **Geometry validity:** neighborhood geometries checked via `is_valid`; repair logic available when needed; invalid geometries are reported.
- **Missingness and duplicates:** records with missing coordinates excluded; duplicate listing IDs filtered (audit saved to `outputs/tables/sample_flow.csv`).
- **Outliers:** hard plausibility thresholds (< 10, $> 10{,}000$) plus winsorization at 0.5% and 99.5% (documented in preprocessing).
- **Weights specification:** kNN weights are built in a projected CRS (EPSG:25830); the selected k is documented and reused consistently across diagnostics and SAR/SEM.
- **Sample audit:** all sample size transitions tracked in `outputs/tables/sample_flow.csv` (raw → cleaned → model sample).
- **Determinism:** fixed random seed (`RANDOM_SEED=42`) used where sampling is applied; scripts avoid non-deterministic row ordering by sorting on stable identifiers when exporting.